INTERVAL MAPPING OF HUMAN QTL USING SIB PAIR DATA

WEN-YUN LI

(Bachelor of Mathematics, East China Normal University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2006

Acknowledgements

I would like to express my gratitude to all those who have helped me to complete this thesis. Without their warmhearted help, this thesis would not have been possible.

First of all, I would like to express my deepest and most sincere gratitude to my supervisor, Associate Professor Zehua Chen. His stimulating guidance and encouragement helped me in all the time of research and writing of this thesis. It was a great pleasure of me to finish this thesis under his supervision.

The help I received from the faculty members, the laboratory staffs and the administrative staffs of the department is gratefully acknowledged. Thanks to Professor Zhidong Bai for his continuous encouragement and timely help. Thanks to Ms Yvonne Chow and Mr Rong Zhang for the assistance with the laboratory work. Thank you all for your support.

I also wish to express my deep gratitude to my friends in this special time. Thanks to Dr Yue Li, Dr Zhen Pang, Ms Ying Hao, Ms Huixia Liu, Ms Rongli Zhang, Mr Yu Liang, Ms Xiuyuan Yan. Thank you for accompanying me, taking care of me and

encouraging me in all these years.

Especially, I would like to give my special thanks to and share this moment of happiness with my parents, my brother and Mr Jian Xiao–my boyfriend. They have rendered me enormous support during the whole tenure of my research.

Contents

Summary

Various regression models based on sib pair data have been developed for mapping quantitative trait loci (QTL) in human since the seminal paper published in 1972 by Haseman and Elston. To which Fulker and Cardon (1994) adapted the idea of interval mapping for increasing the power of QTL mapping. However, in the interval mapping approach of Fulker and Cardon, the statistic for testing QTL effect does not obey the classical statistical theory and hence critical values of the test can not be appropriately determined. In this thesis, we give a unified treatment to all the Haseman-Elston type regression models and propose an alternative approach to interval mapping. A modified Wald test is proposed for the testing of QTL effect. The asymptotic distribution of the modified Wald test statistic is established and hence the critical values or the p-values of the test can be determined. Simulation studies are carried out to verify the validity of the modified Wald test and to demonstrate its desirable power.

Genome wide search is an important area of QTL mapping, and it has been tackled by several authors (Feingold *et al*. 1993, Churchill and Doerge 1994, Rebai *et al*. 1994, 1995, Piepho 2001, Zou *et al*. 2004) in the experimental species. Multiple hypothesis

testing is implicit in the genome search problem, and this makes the control of the overall type I error rate a problem. The key in the genome search problem is to establish certain appropriate threshold that is able to control the overall type I error rate. We propose an alternative test statistic, which, unlike the above mentioned methods, captures the dependence structure of the multiple tests. Method for simulating the thresholds is provided. Simulation studies verify the validity of the test and the power of the test is demonstrated.

The multi-point interval mapping of QTL uses the information carried by more markers rather than only the two flanking markers and is surely more powerful than the two-point interval mapping. The current multi-point interval mapping methods estimate the IBD proportion at the QTL by either linear combination or hidden Markov chain algorithm. In this thesis, we propose an alternative multi-point interval mapping method. We estimate the IBD proportions at the flanking markers with the joint distribution of the numbers of alleles IBD shared at multiple markers, and then perform the two-point interval mapping. This multi-point interval mapping method is shown by simulation study to be more powerful than the two-point interval mapping method under certain situations.

The likelihood ratio (LR) test is always among the most powerful methods. Several researchers have applied the LR test to the interval mapping of QTL (Lander and Botstein 1989, Haley and Knott 1992, Fulker and Cardon 1994, Fulker *et al.* 1995), but none of them have studied the asymptotic distribution of the LR test statistic, which

is not too difficult for the interval mapping problem. We apply the result of Self and Liang (1987) to the interval mapping problem and deduce that the asymptotic distribution of the LR test statistic is a mixture of χ^2_1 and χ^2_2. Simulation studies show that the combination of the LR test and the multi-point interval mapping model possesses the highest power among the 4 combinations of multi-point interval mapping/interval mapping model and the modified Wald/LR test.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to QTL mapping

Many traits in plants, animals, and human beings can be measured on a numerical scale, continuous or discrete, and they are called quantitative traits (QTs). Since many of the QTs have strong genetic determinant and are highly heritable, it is of considerable interest to find the genes underlying such QTs. The process of detecting quantitative trait loci in the genome is called quantitative trait loci (QTL) mapping.

The goals of QTL mapping include: (i) finding the locations in the genome where the QTLs lie in, if exist, (ii) making clear to what extent each QTL influences the QT, and (iii) understanding the structures of the QTLs – their allele frequencies, the contribution of each allele to the QT. Statistical analysis is indispensable in achieving

these goals. The more challenging task of QTL mapping is to achieve the first two goals: mapping the locations and estimating the genetic variances of the QTLs.

An important concept in QTL mapping is the distance between two loci in the genome. Of relevance is the genetic distance instead of the physical distance. The genetic distance is measured by Morgan (or centiMorgan, i.e., a hundredth of a Morgan). One Morgan is defined as the length of the DNA sequence at which exactly one crossover is expected to occur.

The development of early QTL mapping was limited by the lack of densely mapped markers, and the main methods used included ANOVA, linear regression, t test for one-marker cases and F test for multiple-marker cases. In these methods, the markers are thought of as the candidate genes, and so came the name 'candidate gene approach'.

With the advent of Restriction Fragment Length Polymorphism (RFLP) as genetic markers, systematic mapping of QTL became possible in principle (Botstein *et al*. 1980). This gave rise to the development of the 'marker locus approach'. The refinement of statistical methods (Lander and Botstein 1986, 1989) made the marker locus approach very popular. A great deal of the advanced QTL mapping methods in experimental species are based on the idea of interval mapping proposed by Lander and Botstein (1989).

The data used in QTL mapping generally include the quantitative trait values (or simply trait values) and the genotypes at some markers in the vicinity of which the QTL(s) is (are) suspected to locate, and sometimes also include other cofactors affecting

the trait values, such as the environmental factors, the gender of the individual, the pedigree structure, and so on.

1.2 QTL mapping in experimental species and in human

The study on QTL mapping in experimental species is more successful and extensive than that in human. The reason is as follows.

In experimental species, pure homozygous strains(homozygous at every locus) can be generated through selective crossing and can be used for various experimental crosses. For example, let P_1 and P_2 be two parent lines whose genotypes at loci A, B and C are respectively 'ABC/ABC' and 'abc/abc'. The generation produced from the cross between P_1 and P_2 are called F1 generation, whose genotype is 'ABC/abc', heterozygous at every locus. The cross between F1 and one of its parental lines, say P_1, is called a B_1 backcross, and the cross between F1 and F1 is called a F2 intercross. The parental origins of alleles of the offsprings are known unambiguously. This feature renders the testing for equality of the QT mean values in different genotype classes feasible. The environmental variations can also be largely controlled in the experiments. In experimental species, for each individual, the genotype probabilities of an untyped putative QTL flanked by two typed markers can be obtained conditioning on the individual's marker genotypes. Under the assumption that the QT follows a distribution in a known parametric family given the QTL genotypes, a mixture model can be formulated, and

QTL mapping can be done by various methods, for example the maximum likelihood methods or the regression methods.

For human beings, the QT also follows a mixture distribution if the parametric family is assumed. But the mixture structure is much more complicated. In human, an unambiguous identification of parental origins of alleles and control for environmental variations are impossible, because human cannot be bred in controlled crosses and thus no pure inbred lines are available. Therefore, the QTL mapping approaches in experimental species are not applicable to QTL mapping in human.

It is easy to understand that, the more genetic materials two individuals share in common the more similar their QTs are. This is a fundamental idea underlying many approaches to QTL mapping in human. In human QTL mapping, the genetic similarity is represented by the proportion of alleles identical by descent (IBD). Two alleles, which are IBD, are copies of the same allele descended from a common ancestor. Since alleles at linked loci tend to co-segregate, if a pair of relatives share alleles IBD at one locus, they will also share alleles IBD at a linked locus with high probability. Generally, the extent of marker allele IBD sharing is related to the QT similarity. The proportion of alleles IBD will be referred to as 'IBD proportion' in short throughout this thesis. Since siblings share the same parents and in most cases the same living environment, it is easier to analyze the relationship between their QT similarity and IBD proportion than other relative types. Sib pair models play an important role in human QTL mapping.

The calculation of IBD proportion is an important component in sib pair models

and the like for QTL mapping in human. As we know, each person has 2 alleles at each locus, one from the father and the other from the mother, so any two persons can share at most 2 alleles IBD. A general method for calculating the probabilities of sharing 0, 1, and 2 alleles IBD at a locus by a random pair of relatives was given by Li and Sacks (1954). This was then extended by Campbell and Elston (1971), and a more general method was developed by Donnelly (1983).

1.3 Literature review

In this section, approaches for QTL mapping are reviewed. In view of the differences between QTL mapping in experimental species and in human, we will introduce the approaches separately in two subsections.

1.3.1 QTL mapping approaches in experimental species

The availability of dense genetic markers provides the foundation for sophisticated QTL mapping methodologies. These techniques include single marker mapping methods (Edwards *et al*. 1987, Beckmann and Soller 1988, Luo and Kearsey 1989, Simpson 1989, 1992), methods using Bayesian analysis (Hoeschele and VanRaden 1993, Satagopan *et al*. 1996, Uimari and Hoeschele 1997, Sillanpää and Arjas 1999), methods using genetic algorithm (Carlborg *et al*. 2000), interval mapping (Lander and Botstein 1989) and its various extensions: regression based interval mapping (Haley and Knott

1992), composite interval mapping (CIM; Jansen 1993, Zeng 1993, 1994, Jansen and Stam 1994) and multiple interval mapping (MIM; Kao and Zeng 1997, Kao *et al*. 1999, Zeng *et al*. 1999).

There are many excellent reviews on the QTL mapping methods in experimental species (Doerge *et al*. 1997, Liu 1997, Lynch and Walsh 1998, Broman and Speed 1999, Broman 2001, Doerge 2002). In the following, we only give a sketch of major approaches.

The most widely used methods for single marker mapping are based on ANOVA (Soller *et al*. 1976, Edwards *et al*. 1987), *t* test or simple linear regression to assess the segregation of a phenotype with respect to a marker genotype. Though ANOVA at one marker locus can be easily extended to account for multiple loci, it fails to provide an estimate of QTL location.

Thoday (1961) proposed the idea of using two markers to bracket a region for detecting QTL. Lander and Bostein (1989) improved Thoday's idea and proposed the single interval mapping method for experimental organisms. In the single interval mapping method, the QTL effect is estimated at each fixed position in the interval, and thus the QTL effect and QTL location are no longer confounded. The single interval mapping is more powerful than the single marker mapping due to the additional information supplied by the flanking markers. In view of the relative complexity and computational demand of the maximum likelihood estimation used by Lander and Botstein, Haley and Knott (1992) proposed a regression based method to approximate the single interval

mapping method for experimental species. Their method was shown to be asymptotically equivalent to the maximum likelihood based interval mapping of Lander and Botstein (Haley and Knott 1992, Rebai *et al*. 1995).

Quantitative traits are by nature affected by many genes, and thus multiple QTL models are more natural to consider in QTL mapping. In single interval mapping, QTLs are mapped one at a time, ignoring the effects of other QTLs. When multiple QTLs are present, the single interval mapping may yield biased location estimates because of the effects of other QTLs (Lander and Botstein 1989, Haley and Knott 1992, Jansen 1993, Zeng 1994), and it is also less powerful in detecting the QTL. The multiple QTL models, which take into account the effects of multiple QTLs simultaneously, are more efficient and can estimate the QTL locations more accurately (Knapp 1991, Haley and Knott 1992). CIM and MIM are examples of such multiple QTL models.

CIM combines interval mapping with multiple linear regression. Additional markers are included as cofactors to account for the variation associated with other QTLs in the same chromosome and thus the residual variance gets reduced. To detect a QTL Q in the marker interval (M_i, M_{i+1}), the statistical model is generally defined as:

$$y = b_0 + b^* x^* + \sum_{k \neq i,i+1} b_k x_k + e \tag{1.1}$$

for the backcross population, where y is the QT, x^* takes 1 or 0, denoting respectively the homozygous and heterozygous genotype of Q, x_k is a similar genotype indicator for

marker M_k, b^* and b_k denote the effects of Q and M_k respectively; or

$$y = \mu + a^* x^* + d^* z^* + \sum_{k \neq i,i+1} (a_k x_k + d_k z_k) + e \tag{1.2}$$

for the F2 population, where x^* takes 1, -1 or 0 for the two homozygous and one heterozygous genotypes of the QTL respectively, and similarly does x_k for M_k, z^* and z_k are the heterozygous indicators for Q and M_k respectively, and a^*, a_k, d^* and d_k are the corresponding additive and dominant effects. Since the QTL genotypes are unobservable, x^* and z^* in model(1.1) and model(1.2) are missing. Assuming the normality of the random error e, the distribution of y is a mixture of several normal distributions–2 for backcross and 3 for F2, and the mixing probabilities can be determined conditioning on the genotypes of M_i and M_{i+1}. The MLE of the parameters in the above models can be obtained through the EM algorithm. By combining interval mapping with multiple regression, CIM creates a condition that individual QTLs can be separated for testing and estimation.

MIM is an extension of interval mapping to the mapping of multiple QTLs. Multiple marker intervals are used to account for the effects of multiple QTLs. Suppose m intervals are investigated, so there are m putative QTLs if we assume at most one QTL in each interval. The statistical model is defined as:

$$y = \mu + \sum_{r=1}^{m} \alpha_r x_r + e$$

for the backcross population, where y is the QT, x_r takes 1 or 0 for the homozygous and heterozygous genotype of the r-th QTL, Q_r, respectively, and α_r denotes the effect of

the Q_r; or

$$y = \mu + \sum_{r=1}^{m} a_r x_r + \sum_{t=1}^{m} d_t z_t + e$$

for the F2 population, where x_r, z_r, a_r and d_r are defined similarly as in CIM. The QTL genotypes are unobservable, but their probabilities can be analyzed conditioning on the genotypes of the flanking markers of the r-th interval. Assuming the normality of e, the distribution of y is actually a mixture of several normal distributions. The interaction terms of x_rs can also be considered in the two models to account for the epistatic effects. Just like CIM, the EM algorithm can be used to estimate the QTL effects.

For CIM and MIM, when the number of markers under consideration is large, model selection is in order for pinpointing the most appropriate genetic model relating the QT to the QTL (Jansen 1993, Jansen and Stam 1994, Kao *et al*. 1999, Zeng *et al*. 1999).

1.3.2 QTL mapping approaches in human

Haseman-Elston regression is the first statistical method developed for human QTL mapping (Haseman and Elston 1972). This method used sib pair data. The squared difference of sib pair trait values is regressed onto the IBD proportion at a marker. With the advent of dense markers throughout the entire genome, many sophisticated methods for human QTL mapping have been developed based on the idea of Haseman and Elston. The sib pair method also has been extended to other relative pairs and pairs drawn from large pedigrees (Olson and Wijsman 1993).

In the original Haseman-Elston regression, only the information contained in the trait difference is used. Wright (1997) pointed out that the use of trait difference only discards some useful information and suggested to include the squared trait sum in the regression model. Subsequently, Drigalenko (1998) proposed the trait product method, which used the product of the centralized trait values of the sib pair as the response variable. However, the trait product method is only correct in certain situations such as the squared sum and the squared difference have the same variance. To address this problem, a host of approaches called "revised Haseman-Elston" were developed. The "revised Haseman-Elston" approaches use the weighted average of squared difference and squared sum of the sib pair trait values as the response variable. The weights are chosen in such ways that the response and the IBD proportion at the marker are most highly correlated. One such choice is the inverted variances of the squared difference and squared sum (Elston *et al*. 2000, Xu *et al*. 2000, Forrest 2001, Sham and Purcell 2001, Visscher and Hopper 2001). Sham *et al*. (2002) took a further step to extend this method to extended pedigrees. These approaches have achieved great success in terms of power for detecting QTL. Several review papers have devoted to these regression based methods (Feingold 2002, Szatkiewicz *et al*. 2003, Majumder and Ghosh 2005).

In addition to the above mentioned "revised Haseman-Elston" methods, some other competitive methods were also proposed. The variance components models (VC) were proposed by Amos (1994), see also Stern *et al*. (1996), Mitchell *et al*. (1997), Almasy *et al*. (1997), Towne *et al*. (1997) and Almasy and Blangero (1998). The VC models are applicable not only to sib pairs but also to large sibships or pedigrees. The vari-

ance components methods rely heavily on the normality assumption of the traits. When this assumption holds or nearly holds the VC models are very powerful. However, if this assumption is not met, the VC models are poor and can be outperformed by the Haseman-Elston regression methods. The score statistic methods were considered by Tang and Siegmund (2001), Wang and Huang (2002) and Putter *et al*. (2002). The score statistic methods have properties similar to the "revised Haseman-Elston" methods. When due consideration is taken, the score statistic methods are comparable in power with the VC models if the normality assumption holds, and enjoy the robustness of the "revised Haseman-Elston" methods otherwise.

Besides parametric methods, there are also nonparametric methods proposed for QTL mapping in human. For example, the rank based statistic methods were considered by Haseman and Elston (1972), and Kruglyak and Lander (1995), the kernel smoothing methods were considered by Ghosh and Majumder (2000), and Ghosh *et al*. (2003)

Both the original and revised Haseman-Elston regression methods have a common limitation: only the information at one marker is used, and the QTL effect (σ_g^2) and the recombination fraction (θ) between the QTL and the marker cannot be distinguished. As a consequence, the power is low especially when the QTL and the marker are far apart, and only a coarse estimate of the QTL location can be obtained.

Fulker and Cardon (1994) incorporated the idea of interval mapping for experimental species (Lander and Botstein 1989, Haley and Knott 1992), which used two flanking markers of the putative QTL simultaneously rather than one at a time, into the original

Haseman-Elston regression, and proposed the interval mapping method for human QTL mapping. They demonstrated that this method is able to achieve higher power and get more accurate location estimate. However, this method is effective only when the flanking markers are completely informative, that is, the IBD proportions of the flanking markers are known with certainty, as pointed out by Fulker *et al*. (1995). Fulker *et al*. (1995) extended this interval mapping method to a multi-point interval mapping method which uses more than two markers. It has been shown that the multi-point method is effective even when the markers are not completely informative.

1.4 Aim and organization of the thesis

The QTL location estimation in the current interval mapping approaches is accomplished by grid-point searching, which requires either a maximum likelihood estimation or a linear regression at every fixed point in the interval. Furthermore, the search can be multi-dimensional when multiple QTLs present, so the amount of computation is tremendous. In this thesis, we provide a simple and quick approach to QTL location estimation for interval mapping, which requires only one linear regression in each interval.

The *t* test used in the regression based interval mapping of Fulker and Cardon (1994) is not valid due to the inaccurate approximation to the distribution of the test statistic. In this thesis, we provide a modified Wald statistic, whose thresholds can be derived from

the joint distribution function of two correlated standard normal random variables.

In real QTL mapping, the single interval mapping is carried out interval by interval in a genome-wide search manner, and multiple tests are involved in the procedure. Therefore, one needs to determine the unified threshold for controlling the overall Type I error rate. In this thesis, we provide a numerical approximation to this threshold by resampling from a multivariate normal distribution.

A multi-point interval mapping approach is also considered in this thesis. Fulker *et al*. (1995) proposed a multi-point interval mapping approach that estimates the IBD proportion at QTL with a linear combination of IBD proportions at multiple markers. Kruglyak *et al*.(1995) and Lander and Green(1987) suggested the hidden Markov chain approach for multi-point interval mapping that estimates the IBD proportion at QTL using the IBD proportions at multiple markers through the hidden Markov chain algorithm. However, the linear combination expression in the approach of Fulker *et al*. (1995) and the transitional matrices in the hidden Markov chain approach are derived over the entire population and do not take the particular marker genotypes into account. Unlike the above two approaches, our multi-point interval mapping uses the joint probability of the numbers of alleles IBD shared at multiple markers to estimate the IBD proportions at the flanking markers and then performs the single interval mapping. The joint probability of the numbers of alleles IBD at multiple markers is derived by adding up the probabilities of all possible allele-transmission patterns conditioning on the marker genotypes. The estimated IBD proportions at the flanking markers are

marker-genotype specific, and thus should be more accurate than those obtained through the linear combination approach and the hidden Markov chain approach.

Among the current test statistics for interval mapping, none has a closed form asymptotic distribution. In this thesis, we give a closed form asymptotic distribution for the likelihood ratio statistic for interval mapping.

The thesis is organized as follows.

In Chapter 2, we derive the formula for the expected IBD proportion at QTL conditioning on the IBD proportions at the two flanking markers, and then propose a one-step location estimation procedure based on this conditional expectation. Simulation studies are conducted to compare our location estimation procedure with the grid-point searching approach of Fulker and Cardon (1994). A modified Wald test for detecting the QTL effect is then proposed and compared to the ideal t test by a simulation study.

In Chapter 3, a genome-wide search strategy using the modified Wald statistic given in Chapter 2 is proposed. The procedure for simulating the thresholds is outlined. A simulation study is performed to assess the power of this genome-wide search strategy.

In Chapter 4, a new model for multi-point interval mapping is formulated, the procedure for calculating the joint distribution of the numbers of alleles IBD at multiple markers and the multi-point estimates of IBD proportions at flanking markers are described. A simulation study is conducted to compare the single interval mapping using locally estimated IBD proportions at flanking markers and the multi-point interval map-

ping.

In Chapter 5, the likelihood ratio statistic for interval mapping is formulated and then its asymptotic distribution is derived using the result of Self and Liang (1987). A simulation study is performed to compare the likelihood ratio test and the modified Wald test. The influences of the local and multi-point estimates of the IBD proportion on the two tests are also analyzed based on the simulation results.

In Chapter 6, we give the conclusions on the thesis research and discuss some possible directions of further research: the combination of the variance components model with the interval mapping approach, the asymptotic distribution of the likelihood ratio statistic in multiple QTL mapping and the generalized linear model for interval mapping.

Chapter 2

Interval Mapping of QTL in Human

2.1 Haseman-Elston regression model at a fixed locus

The cases considered in the early works in QTL mapping are very simple, in which there is only one QTL responsible for the trait under investigation and no dominant effect of the QTL is assumed. Suppose the QTL under investigation has K different alleles, λ_k denotes the contribution of the k-th allele to the trait value, and p_k denotes the population frequency of the k-th allele. The sib pair trait values can be expressed as,

$$\begin{aligned} x_1 &= \mu + c_{11} + c_{12} + \epsilon_1, \\ x_2 &= \mu + c_{21} + c_{22} + \epsilon_2, \end{aligned} \tag{2.1}$$

where c_{11}, c_{12}, c_{21} and c_{22} are the allele contributions at the QTL, and ϵ_1, ϵ_2 are random errors. c_{11}, c_{12}, c_{21} and c_{22} are independently identically distributed random variables

with $P(c_{ij} = \lambda_k) = p_k,\ k = 1, \cdots, K$.

In the original work of Haseman and Elston, the expectation of the squared sib pair traits difference (Z) conditioning on the IBD proportion at a marker (π_M) was derived as

$$Z = \alpha_D - \beta_{\pi_M}\pi_M + e_D. \tag{2.2}$$

When the QTL is located away from the marker with recombination fraction θ in between,

$$\beta_{\pi_M} = 2(1 - 2\theta)^2\sigma_g^2, \tag{2.3}$$

and when the marker itself is the QTL,

$$\beta_{\pi_M} = 2\sigma_g^2.$$

In cases that π_M cannot be determined unambiguously, replacing π_M with $\hat{\pi}_M$ leads to the same regression model as model (2.2),

$$Z = \alpha_D - \beta_{\hat{\pi}_M}\hat{\pi}_M + e_D, \tag{2.4}$$

where

$$\hat{\pi}_M = f_2 + f_1/2,$$

and f_i (i = 0, 1, or 2) is the probability that the sib pair share i alleles IBD at the marker conditioning on the marker genotypes of the sib pair and their parents. Values of f_i and $\hat{\pi}_M$ can be obtained from Table II of Haseman and Elston (1972).

2.2 Estimation of the proportion of alleles IBD shared at a QTL by a sib pair using the information in flanking markers

An important step in the interval mapping approach we propose in this chapter is to estimate the proportion of alleles IBD shared at a QTL by a sib pair given the proportions of alleles IBD they share at the two flanking markers. In this section, we derive the formula for this estimation through the joint distribution of the IBD proportions at three loci (one QTL and two flanking markers).

2.2.1 Joint distribution of the proportions of alleles IBD shared by a sib pair at three loci

Suppose loci A, B and C are located at alphabetic order on the same autosomal chromosome. Let the recombination fraction between A and B be θ_{AB}, between B and C be θ_{BC}. Assume there is no crossover interference, then the recombination fraction between A and C satisfies: $\theta_{AC} = \theta_{AB}(1 - \theta_{BC}) + (1 - \theta_{AB})\theta_{BC}$.

Consider the mating type of parents at loci A, B and C:

$$\begin{array}{c|c} A_1 & A_2 \\ B_1 & B_2 \\ C_1 & C_2 \end{array} \quad \times \quad \begin{array}{c|c} A_3 & A_4 \\ B_3 & B_4 \\ C_3 & C_4 \end{array}$$

where, the subscripts 1, 2, 3 and 4 denote respectively the origins of the alleles: paternal grandfather, paternal grandmother, maternal grandfather, and maternal grandmother.

For each parental genotype, the 8 possible haplotypes that each parent segregates and their corresponding frequencies are listed in Table 2.1.

Table 2.1: Haplotype frequencies of parents

Parent 1	*Frequency*	*Parent 2*	*Frequency*
(A_1, B_1, C_1)	$(1-\theta_{AB})(1-\theta_{BC})/2$	(A_3, B_3, C_3)	$(1-\theta_{AB})(1-\theta_{BC})/2$
(A_2, B_2, C_2)	$(1-\theta_{AB})(1-\theta_{BC})/2$	(A_4, B_4, C_4)	$(1-\theta_{AB})(1-\theta_{BC})/2$
(A_1, B_1, C_2)	$(1-\theta_{AB})\theta_{BC}/2$	(A_3, B_3, C_4)	$(1-\theta_{AB})\theta_{BC}/2$
(A_2, B_2, C_1)	$(1-\theta_{AB})\theta_{BC}/2$	(A_4, B_4, C_3)	$(1-\theta_{AB})\theta_{BC}/2$
(A_2, B_1, C_1)	$\theta_{AB}(1-\theta_{BC})/2$	(A_4, B_3, C_3)	$\theta_{AB}(1-\theta_{BC})/2$
(A_1, B_2, C_2)	$\theta_{AB}(1-\theta_{BC})/2$	(A_3, B_4, C_4)	$\theta_{AB}(1-\theta_{BC})/2$
(A_1, B_2, C_1)	$\theta_{AB}\theta_{BC}/2$	(A_3, B_4, C_3)	$\theta_{AB}\theta_{BC}/2$
(A_2, B_1, C_2)	$\theta_{AB}\theta_{BC}/2$	(A_4, B_3, C_4)	$\theta_{AB}\theta_{BC}/2$

For simplicity, we introduce some new notations which will be used later as follows:

$$\Psi_{AB} = {\theta_{AB}}^2 + (1-\theta_{AB})^2,$$

$$\Psi_{BC} = {\theta_{BC}}^2 + (1-\theta_{BC})^2,$$

$$\Psi_{AC} = {\theta_{AC}}^2 + (1-\theta_{AC})^2.$$

In this section, all genotypes of siblings are assumed to be phase known, and the origin of each allele is assumed to be known. For each pair of full sibs, define a comparison

vector $v = (i_1, i_2, i_3, i_4, i_5, i_6)$, where $i_k = 0/1$, $k = 1, 2, \cdots, 6$. i_1, i_2 and i_3 indicate respectively whether or not the two alleles of the two sibs inherited from parent 1 at locus A, B and C are IBD (Yes=1, No=0). Similarly, i_4, i_5 and i_6 indicate whether or not the two alleles of the sibs from parent 2 at locus A, B and C are IBD, respectively. Let π_A, π_B and π_C denote the IBD proportion at locus A, B and C, respectively. Obviously, $\pi_A = (i_1 + i_4)/2$, $\pi_B = (i_2 + i_5)/2$ and $\pi_C = (i_3 + i_6)/2$. Given the comparison vector, for any genotype of sib 1, there is one and only one possible genotype for sib 2. Therefore the comparison vector can be used to derive the probability of the genotype of sib 2 from that of sib 1.

The probability of the genotype of one sibling is the product of the frequencies of the two haplotypes inherited from both parents. Except for a constant 1/4 (the probability of inheriting the particular alleles at locus A from both parents), the probability of the genotype of one sibling can be factorized into four factors: (a) the probability of inheriting an allele at locus B given the inherited allele at locus A from parent 1, (b) the probability of inheriting an allele at locus C given the inherited allele at locus B from parent 1, (c) the probability of inheriting an allele at locus B given the inherited allele at locus A from parent 2, (d) the probability of inheriting an allele at locus C given the inherited allele at locus B from parent 2.

Given the comparison vector, each factor of the genotype probability of sib 2 can be deduced from the corresponding factor of sib 1. We now take the first factor as an example to illustrate this process. It can be conceived that the first factor takes only two

values: (1-θ_{AB}) – when the origins (represented by the subscripts) of the alleles at loci A and B inherited from parent 1 are the same, and θ_{AB} – when they are not. To conclude, the first factor of the genotype probability only depends on the equality status of the origins of the two alleles at loci A and B inherited from parent 1, which will be referred to as "equality status" for simplicity. If both origins of the alleles at A and B of sib 2 are the same as those of sib 1 ($i_1 = i_2 = 1$) or both are different from those of sib 1 ($i_1 = i_2 = 0$), that is $i_1 = i_2$, the equality status at loci A and B are the same for sib 1 and sib 2, and thus the first factor of the genotype probability of sib 1 and sib 2 are equal. For example, if sib 1 inherits A_1B_1 from parent 1 and $i_1 = i_2 = 0$, then the equality status for sib 1 is "same origin" and the first factor of the genotype probability of sib 1 is (1-θ_{AB}), the equality status at A and B for sib 2 should also be "same origin" since $i_1 = i_2 = 0$, and thus the first factor of the genotype probability of sib 2 is also (1-θ_{AB}). For the above example, we can also deduce that sib 2 inherits A_2B_2 from parent 1 and the first factor of its genotype probability is (1-θ_{AB}), the result remains the same. On the other hand, if one and only one of the origins of the 2 alleles at loci A and B inherited from parent 1 of sib 2 is the same as that of sib 1, that is $i_1 \neq i_2$, the equality status will be different between sib 1 and sib 2 and so are the first factors of the genotype probability of the sib pair, and thus the first factor of the genotype probability of one sib is $(1 - \theta_{AB})$ and the other must be θ_{AB}. For example, if sib 1 inherits A_1B_1 and ($i_1 = 1$, $i_2 = 0$), then sib 2 inherits A_1B_2, their first factors of the genotype probability are (1-θ_{AB}) and θ_{AB}, respectively. The relationships of other factors of the genotype probability between sib 1 and sib 2 are similar.

For any given value (a, b, c), the probability P($\pi_A = a$, $\pi_B = b$, $\pi_C = c$) equals the total probability of all possible sib pair genotypes with IBD proportions a, b and c at loci A, B and C, respectively. In this regard, the specific genotypes of the sib pair are not essential, and all genotypes with the same probability can be combined to form one group. It can be found from Table 2.1 that the 8 haplotypes transmitted by one parent can be classified to 4 groups according to their frequencies, and each group contains two haplotypes. Therefore, each possible genotype probability of one sibling corresponds to a group of 4 genotypes, and there are totally 16 such groups. For example, the group of genotypes corresponding to probability $(1-\theta_{AB})^2\theta_{BC}(1-\theta_{BC})/4$ contains $A_1B_1C_2/A_3B_3C_3$, $A_1B_1C_2/A_4B_4C_4$, $A_2B_2C_1/A_3B_3C_3$ and $A_2B_2C_1/A_4B_4C_4$. For each group of 4 genotypes of sib 1, the corresponding 4 genotypes of sib 2 satisfying the given comparison vector must have the same probability and thus are in the same group.

The detailed computation procedure can be illustrated by examples.

Consider a sib pairs with $\pi_A = 0$, $\pi_B = 1$, $\pi_C = 0$. There is only one possible comparison vector: (0,1,0,0,1,0), and $i_1 \neq i_2$,$i_2 \neq i_3$,$i_4 \neq i_5$,$i_5 \neq i_6$. Therefore whatever is the genotype of sib 1, all 4 factors of the genotype probability of sib 2 are different from those of sib 1. Therefore the probability of such a sib pair is

$$\frac{1}{16}{\theta_{AB}}^2(1-\theta_{AB})^2{\theta_{BC}}^2(1-\theta_{BC})^2.$$

Since there are 64 such pairs of genotypes, thus

$$\begin{aligned} P((0,1,0)) &= 64 \cdot {\theta_{AB}}^2(1-\theta_{AB})^2{\theta_{BC}}^2(1-\theta_{BC})^2/16 \\ &= (1-\Psi_{AB})^2(1-\Psi_{BC})^2/4. \end{aligned}$$

A more complicated case: $\pi_A = 0, \pi_B = 1/2, \pi_C = 0$, there are 2 possible comparison vectors: (0,0,0,0,1,0) and (0,1,0,0,0,0). For the first comparison vector, $i_1 = i_2, i_2 = i_3$, the first two factors of the genotype probability are the same for the pair of sibs; $i_4 \neq i_5, i_5 \neq i_6$, thus the last 2 factors differ between the 2 sibs. So the probability of the first comparison vector is:

$$\begin{aligned} & 16 \cdot [(1-\theta_{AB})^2(1-\theta_{BC})^2 + (1-\theta_{AB})^2{\theta_{BC}}^2 + {\theta_{AB}}^2(1-\theta_{BC})^2 + {\theta_{AB}}^2{\theta_{BC}}^2] \\ & \quad \cdot [\theta_{AB}(1-\theta_{AB})\theta_{BC}(1-\theta_{BC})]/16 \\ = \; & \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})/4. \end{aligned}$$

The same result can be obtained for the second comparison vector. Thus,

$$P((0,1/2,0)) = \frac{1}{2}\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC}).$$

The joint probabilities of π_A, π_B and π_C when $\pi_C = 0$ are:

$$
\begin{aligned}
P((0,0,0)) &= \frac{1}{4}{\Psi_{AB}}^2{\Psi_{BC}}^2, \\
P((1,0,0)) &= (1-\Psi_{AB})^2{\Psi_{BC}}^2/4, \\
P((0,1,0)) &= (1-\Psi_{AB})^2(1-\Psi_{BC})^2/4, \\
P((1,1,0)) &= {\Psi_{AB}}^2(1-\Psi_{BC})^2/4, \\
P((0,1/2,0)) &= \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})/2, \\
P((1,1/2,0)) &= \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})/2, \\
P((1/2,0,0)) &= P(\pi_B = \pi_C = 0) - P((0,0,0)) - P((1,0,0)), \\
&= \Psi_{AB}(1-\Psi_{AB}){\Psi_{BC}}^2/2, \\
P((1/2,1,0)) &= P(\pi_B = 1, \pi_C = 0) - P((0,1,0)) - P((1,1,0)), \\
&= \Psi_{AB}(1-\Psi_{AB})(1-\Psi_{BC})^2/2, \\
P((1/2,1/2,0)) &= P(\pi_C = 0) - P((0,0,0)) - P((1,0,0)) - P((0,1,0)) - P((1,1,0)), \\
&\quad -P((0,1/2,0)) - P((1,1/2,0)) - P((1/2,0,0)) - P((1/2,1,0)), \\
&= ({\Psi_{AB}}^2 + (1-\Psi_{AB})^2)\Psi_{BC}(1-\Psi_{Bc})/2.
\end{aligned}
$$

When $\pi_C = 1$, the following equations can be verified:

$$
\begin{aligned}
P((0,0,1)) &= P((1,1,0)) = {\Psi_{AB}}^2(1-\Psi_{BC})^2/4,\\
P((1,0,1)) &= P((0,1,0)) = (1-\Psi_{AB})^2(1-\Psi_{BC})^2/4,\\
P((0,1,1)) &= P((1,0,0)) = (1-\Psi_{AB})^2{\Psi_{BC}}^2/4,\\
P((1,1,1)) &= P((0,0,0)) = {\Psi_{AB}}^2{\Psi_{BC}}^2/4,\\
P((0,1/2,1)) &= P((1,1/2,0)) = \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})/2,\\
P((1,1/2,1)) &= P((0,1/2,0)) = \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})/2,\\
P((1/2,0,1)) &= P((1/2,1,0)) = \Psi_{AB}(1-\Psi_{AB})(1-\Psi_{BC})^2/2,\\
P((1/2,1,1)) &= P((1/2,0,0)) = \Psi_{AB}(1-\Psi_{AB}){\Psi_{BC}}^2/2,\\
P((1/2,1/2,1)) &= P((1/2,1/2,0)) = ({\Psi_{AB}}^2+(1-\Psi_{AB})^2)\Psi_{BC}(1-\Psi_{Bc})/2.
\end{aligned}
$$

The joint probabilities when $\pi_C = 1/2$ are as follows,

$$\begin{aligned}
P((0,0,1/2)) &= {\Psi_{AB}}^2\Psi_{BC}(1-\Psi_{BC})/2,\\
P((0,1/2,1/2)) &= \Psi_{AB}(1-\Psi_{AB})({\Psi_{BC}}^2+(1-\Psi_{BC})^2)/2,\\
P((0,1,1/2)) &= (1-\Psi_{AB})^2\Psi_{BC}(1-\Psi_{BC})/2,\\
P((1/2,0,1/2)) &= \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC}),\\
P((1/2,1/2,1/2)) &= ({\Psi_{AB}}^2+(1-\Psi_{AB})^2)({\Psi_{BC}}^2+(1-\Psi_{BC})^2)/2,\\
P((1/2,1,1/2)) &= P((1/2,0,1/2)) = \Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC}),\\
P((1,0,1/2)) &= P((0,1,1/2)) = (1-\Psi_{AB})^2\Psi_{BC}(1-\Psi_{BC})/2,\\
P((1,1/2,1/2)) &= P((0,1/2,1/2)) = \Psi_{AB}(1-\Psi_{AB})({\Psi_{BC}}^2+(1-\Psi_{BC})^2)/2,\\
P((1,1,1/2)) &= P((0,0,1/2)) = {\Psi_{AB}}^2\Psi_{BC}(1-\Psi_{BC})/2.
\end{aligned}$$

In the next subsection, the conditional expectation of π_B conditioning on π_A and π_C will be derived from the joint density of π_A, π_B and π_C obtained in this subsection.

2.2.2 Estimation of the proportion of alleles IBD shared at a QTL by a sib pair using information in flanking markers

To obtain the conditional expectation of π_B given (π_A, π_C), we need first find out the conditional distribution of π_B. Its values are calculated and listed in Table 2.2.

Table 2.2: Conditional probabilities of π_B given (π_A, π_C)

	π_B		
(π_A, π_C)	0	$\frac{1}{2}$	1
(0,0)	$\frac{\Psi_{AB}{}^2\Psi_{BC}{}^2}{\Psi_{AC}{}^2}$	$\frac{2\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}{}^2}$	$\frac{(1-\Psi_{AB})^2(1-\Psi_{BC})^2}{\Psi_{AC}{}^2}$
$(0,\frac{1}{2})$	$\frac{\Psi_{AB}{}^2\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{\Psi_{AB}(1-\Psi_{AB})(\Psi_{BC}{}^2+(1-\Psi_{BC})^2)}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{(1-\Psi_{AB})^2\Psi_{BC}(1-\Psi_{Bc})}{\Psi_{AC}(1-\Psi_{AC})}$
(0,1)	$\frac{\Psi_{AB}{}^2(1-\Psi_{BC})^2}{(1-\Psi_{AC})^2}$	$\frac{2\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})}{(1-\Psi_{Ac})^2}$	$\frac{(1-\Psi_{AB})^2\Psi_{BC}{}^2}{(1-\Psi_{Ac})^2}$
$(\frac{1}{2},0)$	$\frac{\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}{}^2}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{(\Psi_{AB}{}^2+(1-\Psi_{AB})^2)\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{\Psi_{AB}(1-\Psi_{AB})(1-\Psi_{BC})^2}{\Psi_{AC}(1-\Psi_{AC})}$
$(\frac{1}{2},\frac{1}{2})$	$\frac{2\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}{}^2+(1-\Psi_{AC})^2}$	$\frac{(\Psi_{AB}{}^2+(1-\Psi_{AB})^2)(\Psi_{BC}{}^2+(1-\Psi_{BC})^2)}{\Psi_{AC}{}^2+(1-\Psi_{AC})^2}$	$\frac{2\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}{}^2+(1-\Psi_{AC})^2}$
$(\frac{1}{2},1)$	$\frac{\Psi_{AB}(1-\Psi_{AB})(1-\Psi_{BC})^2}{\Psi_{Ac}(1-\Psi_{AC})}$	$\frac{(\Psi_{AB}{}^2+(1-\Psi_{AB})^2)\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}{}^2}{\Psi_{AC}(1-\Psi_{AC})}$
(1,0)	$\frac{(1-\Psi_{AB})^2\Psi_{BC}{}^2}{(1-\Psi_{AC})^2}$	$\frac{2\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})}{(1-\Psi_{AC})^2}$	$\frac{\Psi_{AB}{}^2(1-\Psi_{BC})^2}{(1-\Psi_{AC})^2}$
$(1,\frac{1}{2})$	$\frac{(1-\Psi_{AB})^2\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{\Psi_{AB}(1-\Psi_{AB})(\Psi_{BC}{}^2+(1-\Psi_{BC})^2)}{\Psi_{AC}(1-\Psi_{AC})}$	$\frac{\Psi_{AB}{}^2\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}(1-\Psi_{AC})}$
(1,1)	$\frac{(1-\Psi_{AB})^2(1-\Psi_{BC})^2}{\Psi_{Ac}{}^2}$	$\frac{2\Psi_{AB}(1-\Psi_{AB})\Psi_{BC}(1-\Psi_{BC})}{\Psi_{AC}{}^2}$	$\frac{\Psi_{AB}{}^2\Psi_{BC}{}^2}{\Psi_{AC}{}^2}$

The next step is to compute the conditional expectations of π_B. In this step, two equations can be used to simplify the formulae. Recall

$$P(\pi_A = 0, \pi_B = 1/2, \pi_C = 0) = \Psi_{AB}(1 - \Psi_{AB})\Psi_{BC}(1 - \Psi_{BC})/2,$$

which also equals

$$\begin{aligned} & P(\pi_A = \pi_C = 0) - P(\pi_A = \pi_B = \pi_C = 0) - P(\pi_A = 0, \pi_B = 1, \pi_C = 0) \\ & = {\Psi_{AC}}^2/4 - {\Psi_{AB}}^2{\Psi_{BC}}^2/4 - (1 - \Psi_{AB})^2(1 - \Psi_{BC})^2/4, \end{aligned}$$

thus,

$$\Psi_{AC} = \Psi_{AB}\Psi_{BC} + (1 - \Psi_{AB})(1 - \Psi_{BC}).$$

Moreover

$$\begin{aligned} & P(\pi_A = 1, \pi_B = 1/2, \pi_C = 0) \\ & = \Psi_{AB}(1 - \Psi_{AB})\Psi_{BC}(1 - \Psi_{BC})/2 \\ & = P(\pi_A = 1, \pi_C = 0) - P(\pi_A = 1, \pi_B = \pi_C = 0) - P(\pi_A = \pi_B = 1, \pi_C = 0) \\ & = (1 - \Psi_{AC})^2/4 - (1 - \Psi_{AB})^2{\Psi_{BC}}^2/4 - {\Psi_{AB}}^2(1 - \Psi_{BC})^2/4, \end{aligned}$$

and thus,

$$1 - \Psi_{AC} = (1 - \Psi_{AB})\Psi_{BC} + \Psi_{AB}(1 - \Psi_{BC}).$$

The conditional expectations of π_B are listed in Table 2.3.

Using Table 2.3, a general formula for the conditional expectations can be derived as:

$$\begin{aligned} E(\pi_B \mid \pi_A, \pi_C) \quad = \quad & \frac{(1 - \Psi_{AB})(1 - \Psi_{BC})}{\Psi_{AC}} + \pi_A \left[\frac{\Psi_{AB}(1 - \Psi_{BC})}{1 - \Psi_{AC}} - \frac{(1 - \Psi_{AB})(1 - \Psi_{BC})}{\Psi_{AC}} \right] \\ & + \pi_C \left[\frac{(1 - \Psi_{AB})\Psi_{BC}}{1 - \Psi_{AC}} - \frac{(1 - \Psi_{AB})(1 - \Psi_{BC})}{\Psi_{AC}} \right]. \end{aligned} \tag{2.5}$$

Table 2.3: Conditional expectations of π_B given (π_A, π_C)

	π_A		
π_C	0	$\frac{1}{2}$	1
0	$\frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}}$	$\frac{(1-\Psi_{AB})(1-\Psi_{BC})}{2\Psi_{AC}} + \frac{\Psi_{AB}(1-\Psi_{BC})}{2(1-\Psi_{AC})}$	$\frac{\Psi_{AB}(1-\Psi_{BC})}{1-\Psi_{AC}}$
$\frac{1}{2}$	$\frac{(1-\Psi_{AB})(1-\Psi_{BC})}{2\Psi_{AC}} + \frac{(1-\Psi_{AB})\Psi_{BC}}{2(1-\Psi_{AC})}$	1/2	$\frac{\Psi_{AB}\Psi_{BC}}{2\Psi_{AC}} + \frac{\Psi_{AB}(1-\Psi_{BC})}{2(1-\Psi_{AC})}$
1	$\frac{(1-\Psi_{AB})\Psi_{BC}}{1-\Psi_{AC}}$	$\frac{\Psi_{AB}\Psi_{BC}}{2\Psi_{AC}} + \frac{(1-\Psi_{AB})\Psi_{BC}}{2(1-\Psi_{AC})}$	$\frac{\Psi_{AB}\Psi_{BC}}{\Psi_{AC}}$

In 2.4.1, the coefficients of π_A and π_C in formula (2.5) will be proved to be identical to those obtained by Fulker and Cardon (1994), which will be introduced in detail in the next section.

2.3 Interval mapping

In the Haseman-Elston regression mentioned in Section 2.1, the genetic variance σ_g^2 and the recombination fraction θ are confounded in the regression line slope β_{π_M}. Furthermore, if the marker is located far away from the QTL (θ is close to 1/2), even if σ_g^2 is large, the power of detecting the QTL effect could be very low. With the advent of dense genome maps of genetic markers, more markers can be used simultaneously in the regression models, and the above problem can thus be avoided.

Based on Thoday's idea (1961) of using two markers to bracket a region for de-

tecting the QTL, Lander and Botstein (1989) proposed the interval mapping method for the experimental species. They used the maximum likelihood estimation and constructed a profile of LOD scores to detect and locate the QTL. They applied this method to the backcross data in their simulation study. In view of the relative complexity and demanding computation of the LOD score method, Haley and Knott (1992) proposed a regression based interval mapping method for experimental species and applied it to the F2 data in the simulation study. They constructed a score profile with the value of $n \ln(RSS_{reduced}/RSS_{full})$, which they proved can provide very close approximation to the likelihood ratio statistic. Haley and Knott (1992) proved by simulation study that their regression based interval mapping is asymptotically equivalent to the maximum likelihood based method of Lander and Botstein, and this was also confirmed by Rebai *et al*. (1995).

To tackle the problems found in Haseman and Elston's single marker mapping approach for human QTL, Fulker and Cardon (1994) extended Haley and Knott's (1992) regression based interval mapping method to the Haseman-Elston regression model. Fulker and Cardon's method will be described in Section 2.3.1 in detail.

2.3.1 Fulker and Cardon's approach and its limitations

Fulker and Cardon extended the regression based interval mapping approach, which was proposed by Haley and Knott (1992) for experimental species, to interval mapping of human QTL. They proceeded as follows.

1. Divide the interval by equally spaced dense grid points,and each point is taken as a putative location of the QTL.

2. At each putative location d, $\pi_q(d)$ is estimated for each sib pair using their genotypes at the two flanking markers simultaneously, rather than one at a time.

3. The squared difference of sib pair trait values is regressed onto $\hat{\pi}_q(d)$.

4. The location d that achieves the smallest residual sum of squares among all the grid points is taken as the estimated location of the QTL and the usual t statistic evaluated at this location is used to test the significance of the QTL effect.

The step 2 – estimate the IBD proportion at each putative QTL using the genotypes at two flanking markers – is implemented by the following rational.

Suppose an interval flanked by two markers, M_1 and M_2, is of concern and the length of the interval is γ in terms of recombination fraction. Let r (unknown) be the recombination fraction between M_1 and a putative QTL, Q, in the interval. Denote by s the recombination fraction between Q and M_2. All these recombination fractions can be calculated with Haldane's mapping function. Assume that there is no crossover interference and hence

$$\gamma = r + s - 2rs. \tag{2.6}$$

Let $\pi_q(d)$ be the proportion of alleles IBD shared at the putative location d by a sib pair. An estimate of $\pi_q(d)$ can be obtained using the marker information of the sib pair, which only depends on the two markers and the putative location d. The estimate of $\pi_q(d)$ is a

linear combination of π_1 and π_2 (the IBD proportions at M_1 and M_2),

$$\hat{\pi}_q(d) = \alpha_M(d) + \beta_{M1}(d)\pi_1 + \beta_{M2}(d)\pi_2,$$

where the coefficients $\beta_{M1}(d)$ and $\beta_{M2}(d)$ can be solved by the normal equations

$$\begin{pmatrix} \mathrm{Cov}(\pi_1, \hat{\pi}_q(d)) \\ \mathrm{Cov}(\pi_2, \hat{\pi}_q(d)) \end{pmatrix} = \begin{pmatrix} \mathrm{V}(\pi_1) & \mathrm{Cov}(\pi_1, \pi_2) \\ \mathrm{Cov}(\pi_1, \pi_2) & \mathrm{V}(\pi_2) \end{pmatrix} \begin{pmatrix} \beta_{M1}(d) \\ \beta_{M2}(d) \end{pmatrix}.$$

Substituting the following results (Elston and Keats 1985, SAGE 1989): $\mathrm{V}(\pi) = 1/8$ and $\mathrm{Cov}(\pi_1, \pi_2) = (1-2\gamma)^2/8$, yields the form of $\beta_{M1}(d)$, $\beta_{M2}(d)$ and $\alpha_M(d)$,

$$\begin{aligned} \beta_{M1}(d) &= [(1-2r(d))^2 - (1-2s(d))^2(1-2\gamma)^2]/[1-(1-2\gamma)^4], \\ \beta_{M2}(d) &= [(1-2s(d))^2 - (1-2r(d))^2(1-2\gamma)^2]/[1-(1-2\gamma)^4], \\ \alpha_M(d) &= [1-\beta_{M1}(d) - \beta_{M2}(d)]/2. \end{aligned}$$

Fulker and Cardon's interval mapping approach is more powerful than the original Haseman-Elston method, and it can also provide a QTL location estimate. Yet, there are still certain problems remaining in the interval mapping approach of Fulker and Cardon. The salient one is that the asymptotic distribution of the t statistic evaluated at the estimated QTL location is unknown. In fact, this statistic arises from the minimization of the residual sum of squares over the whole interval, and its distribution is quite complicated. Based on a limited simulation study (with simulation size 400), Fulker and Cardon (1994) claimed that this statistic conformed approximately to a t distribution. However this claim is doubtful. In the simulation study of this chapter, no evidence

is found to support this claim. Even in their own simulation results, there are certain discrepancies between the nominal t critical values and the simulated critical values.

2.3.2 A unified interval mapping regression model with sib pair data

Let (X_{1i}, X_{2i}, G_i), $i = 1, \ldots, n$, be the observations on n sib pairs, where X_{1i} and X_{2i} are respectively the trait values of the first and the second sib in the i-th sib pair, and G_i is the overall genotype information of the sib pair. The contents of G_i include but are not confined to the genotypes of the sib pair and their parents at the two flanking markers. Denote by Y_i^D the squared difference $[X_{1i} - X_{2i}]^2$ and Y_i^S the centered squared sum $[(X_{1i} - \mu_X) + (X_{2i} - \mu_X)]^2$, where μ_X is the population mean of the trait value.

Fulker and Cardon's interval mapping approach uses only the information contained in Y_i^D. However, using trait difference only discards some useful information (Wright 1997). This problem can be addressed by using the weighted average of Y_i^D and Y_i^S as the response variable, and this is exactly the foundation of the "revised Haseman-Elston" approaches. In 2.4.2, it is showed that the slopes of the model regressing Y^D against π_q and the model regressing Y^S against π_q have the same magnitude but opposite sign. Let ω be a positive number between 0 and 1, then the regression model with response variable given by $Z_i(\omega) = \omega Y_i^S - (1 - \omega) Y_i^D$ will have the same slope as Y^S,

not affected by ω. The regression model is of the form:

$$Z_i(\omega) = E[Z_i(\omega)|G_i] + e_i,$$

where $E[Z_i(\omega)|G_i]$ is the conditional expectation of $Z_i(\omega)$ given G_i. This general form includes all the regression models mentioned in section 1.3.2 as special cases. For example, when $\omega = 0$ and G_i contains only the information on a single marker, this reduces to the original Haseman-Elston regression model. Since we are only interested in comparing our approaches with some existing approaches, rather than the advantage of using combined Y_i^D and Y_i^S as response variable, only the special case of $Z_i(\omega)$ when $\omega = 0$ (the negative squared difference of the sib pair trait values) is taken as the response variable Z for all the simulation studies in this thesis.

Assume that

$$\begin{cases} X_{1i} = g_{1i} + \epsilon_{1i}, \\ X_{2i} = g_{2i} + \epsilon_{2i}, \end{cases}$$

where $g_{1i} = c_{11i} + c_{12i}$ and $g_{2i} = c_{21i} + c_{22i}$ are the genotypic values that depend on the genotypes of the QTL, and ϵ_{1i} and ϵ_{2i} are random errors.

Let σ_g^2 denote the genetic variance, i.e., the common variance of g_{1i} and g_{2i}, σ_ϵ^2 the common variance of ϵ_{1i} and ϵ_{2i}, and ρ the correlation coefficient between ϵ_{1i} and ϵ_{2i}. The genetic variance σ_g^2 is decomposed as an additive component σ_a^2 plus a dominant component σ_d^2. Without loss of generality, we assume that $\sigma_d^2 = 0$ and hence $\sigma_g^2 = \sigma_a^2$. Let π_{qi}, π_{1i} and π_{2i} denote the proportions of alleles IBD shared by the i-th sib pair at Q,

M_1 and M_2, respectively. Using the formula $E[E(X|Y)] = E[X]$, it can be derived that

$$E[Z_i(\omega)|G_i] = E[E(Z_i(\omega)|\pi_{qi})|G_i],$$

$$E[\pi_{qi}|G_i] = E[E(\pi_{qi}|\pi_{1i}, \pi_{2i})|G_i].$$

In 2.4.2, we show that

$$E(Z_i(\omega)|\pi_{qi}) = \alpha_Q(\omega) + \beta_Q \pi_{qi},$$

where $\alpha_Q(\omega) = 2(\sigma_g^2 + \sigma_\epsilon^2)(2\omega - 1) + 2\rho\,\sigma_\epsilon^2$, and $\beta_Q = 2\sigma_g^2$. Note that only $\alpha_Q(\omega)$ is affected by ω. Hence

$$E[Z_i(\omega)|G_i] = \alpha_Q(\omega) + \beta_Q E[\pi_{qi}|G_i].$$

Furthermore, we have

$$E(\pi_{qi}|\pi_{1i}, \pi_{2i}) = \alpha_M + \beta_{M1}\pi_{1i} + \beta_{M2}\pi_{2i},$$

where

$$\begin{aligned}
\beta_{M1} &= \frac{(1-2r)^2 - (1-2s)^2(1-2\gamma)^2}{[1-(1-2\gamma)^4]}, \\
\beta_{M2} &= \frac{(1-2s)^2 - (1-2r)^2(1-2\gamma)^2}{[1-(1-2\gamma)^4]}, \\
\alpha_M &= \frac{1}{2}(1 - \beta_{M1} - \beta_{M2}),
\end{aligned}$$

then

$$E[\pi_{qi}|G_i] = \alpha_M + \beta_{M1}E(\pi_{1i}|G_i) + \beta_{M2}E(\pi_{2i}|G_i).$$

Eventually we arrive at the regression model,

$$Z_i(\omega) = \alpha + \beta_1 E(\pi_{1i}|G_i) + \beta_2 E(\pi_{2i}|G_i) + e_i, \tag{2.7}$$

where $e_i \sim \mathrm{N}(0, \sigma_e^2)$, $\alpha = \alpha_Q(\omega) + \beta_Q \alpha_M$, $\beta_1 = 2\sigma_g^2 \beta_{M1}$ and $\beta_2 = 2\sigma_g^2 \beta_{M2}$. Note that, if G_i is the information on the parents' and the sib pair's genotypes at the two markers, the two conditional expectations in model (2.7) can be obtained from Table II of Haseman and Elston (1972).

Since the interval length is assumed known and fixed, only one parameter is necessary for representing the putative QTL location. From equation (2.6), we have

$$1 - 2\gamma = (1 - 2r)(1 - 2s),$$

and thus β_{M1} and β_{M2} can be transformed as

$$\begin{aligned} \beta_{M1} &= \frac{(1-2r)^4 - (1-2\gamma)^4}{(1-2r)^2[1-(1-2\gamma)^4]}, \\ \beta_{M2} &= \frac{(1-2\gamma)^2[1-(1-2r)^4]}{(1-2r)^2[1-(1-2\gamma)^4]}. \end{aligned}$$

Fulker *et al*. (1995) considered an extension of the interval mapping approach of Fulker and Cardon (1994) that takes into account the information in multiple markers. In what they termed as "the multi-point interval mapping", they proceeded as follows. Let the markers in the same chromosome be ordered from left to right. Divide the interval from the leftmost marker to the rightmost marker by equally spaced grid points. Each grid point is taken as a putative location of the QTL, and the IBD proportion at the putative QTL is estimated with the IBD proportions at all the markers. The squared difference of the sib pair trait values is regressed onto the estimated IBD proportion at the putative QTL as in the single interval mapping. However, the same problems associated with single interval mapping as we pointed out in 2.3.1 linger in the multi-point interval mapping.

The general regression model (2.7) we just derived provides an alternative approach to tackling the multi-point interval mapping. In the case of multiple markers, G_i will represent the information in all these markers. Instead of estimating the IBD proportion at the putative QTL, the two conditional expectations of IBD proportions at the flanking markers in model (2.7) are estimated using the information in all the markers. Thus the single and the multi-point interval mapping can be treated in a unified manner. The multi-point interval mapping will be elaborated in Chapter 4.

It is interesting to notice that there is a one-to-one correspondence between (α_Q, β_Q, r) and $(\alpha, \beta_1, \beta_2)$ when $\sigma_g^2 \neq 0$. Both of these two sets of parameters can be used to parameterize the regression model. But there is a profound difference between the two parameterizations. The parametrization using (α_Q, β_Q, r) is not identifiable when $\sigma_g^2 = 0$. In contrast, this problem does not arise with $(\alpha, \beta_1, \beta_2)$, since the parameterization using $(\alpha, \beta_1, \beta_2)$ is always identifiable.

2.3.3 A one-step estimation procedure

In the formulation of Fulker and Cardon (1994), the regression model is expressed as

$$Y_i^D = \alpha_Q + \beta_Q(\alpha_M + \beta_{M1}\hat{\pi}_{1i} + \beta_{M2}\hat{\pi}_{2i}) + e_i, \tag{2.8}$$

where $e_i \sim \mathrm{N}(0, \sigma_e^2)$ Note that, for fixed r, the coefficients α_M, β_{M1}, and β_{M2} are completely determined. In the estimation procedure of Fulker and Cardon, for each fixed r ranging from 0 to γ, model (2.8) is fitted by minimizing the sum of squares $\sum_{i=1}^{n} e_i^2$

with respect to α_Q and β_Q. The final estimates of α_Q, β_Q and r are the fitted values corresponding to the minimum of the residual sum of squares. This procedure is equivalent to minimizing $\sum_{i=1}^{n} e_i^2$ simultaneously with respect to α_Q, β_Q and r. Because of the one-to-one correspondence between (α_Q, β_Q, r) and $(\alpha, \beta_1, \beta_2)$, the minimization in turn amounts to minimizing $\sum_{i=1}^{n} e_i^2$ simultaneously with respect to α, β_1 and β_2. The estimation procedure is then boiled down to the least squares estimation of model (2.7). This gives rise to the one-step estimation procedure.

In the one-step estimation procedure, $\sum_{i=1}^{n} e_i^2$ is minimized first to obtain the least squares estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$. The estimate of the QTL location, $\hat{r}$, is then obtained by solving the following equations:

$$\begin{aligned} \beta_Q \beta_{M1} &= \frac{\beta_Q[(1-2r)^4 - (1-2\gamma)^4]}{(1-2r)^2[1-(1-2\gamma)^4]} = \hat{\beta}_1, \\ \beta_Q \beta_{M2} &= \frac{\beta_Q(1-2\gamma)^2[1-(1-2r)^4]}{(1-2r)^2[1-(1-2\gamma)^4]} = \hat{\beta}_2, \end{aligned}$$

with the adjustment that whenever $\hat{\beta}_1$ or $\hat{\beta}_2$ is less than zero it is reset to zero in the equations. We are not concerned with the estimation of β_Q although it can be obtained by solving the above equations as well, since our test for the significance of the QTL effect will be based on $\hat{\beta}_1$ and $\hat{\beta}_2$ rather than on $\hat{\beta}_Q$. We have the following results:

$$\hat{r} = \begin{cases} \frac{1}{2}\left[1 - \left(\frac{(\hat{\beta}_1/\hat{\beta}_2)(1-2\gamma)^2 + (1-2\gamma)^4}{1+(\hat{\beta}_1/\hat{\beta}_2)(1-2\gamma)^2}\right)^{1/4}\right], & \text{if } \hat{\beta}_1 > 0, \hat{\beta}_2 > 0; \\ \gamma, & \text{if } \hat{\beta}_1 < 0, \hat{\beta}_2 > 0; \\ 0, & \text{otherwise.} \end{cases}$$

In fact, in the case that both $\hat{\beta}_1$ and $\hat{\beta}_2$ are less than zero, the estimate of r, which is irrelevant, can take any number between 0 and γ.

When the values of the regression coefficients that minimize $\sum_{i=1}^{n} e_i^2$ fall into their allowable range, i.e., $\beta_1 \geq 0$ and $\beta_2 \geq 0$, the one-step procedure can produce the same estimate of r as Fulker and Cardon's procedure if the grid-points are fine enough. When the minimizing values are outside of the allowable range, the estimate of r yielded by the one-step procedure is either 0 or γ while that yielded by Fulker and Cardon's procedure is usually between 0 and γ. However the difference is just little.

2.3.4 A modified Wald test

Before we discuss the modified Wald test, let us take another look at the test statistic used by Fulker and Cardon. Their statistic is given by

$$t(\hat{r}) = \frac{\hat{\beta}_Q(\hat{r})}{\hat{\sigma}(\hat{\beta}_Q(\hat{r}))},$$

where $\hat{r}$ is the estimate of r at which the minimum of the residual sum of squares is attained, and $t(\hat{r})$ is the usual t statistic when $\hat{r}$ is taken as if it is a fixed value. Some straightforward algebra in 2.4.3 shows that

$$[t(\hat{r})]^2 = \max_{0 \leq r \leq \gamma} [t(r)]^2 = \max_{0 \leq r \leq \gamma} \left[\frac{\hat{\beta}_Q(r)}{\hat{\sigma}(\hat{\beta}_Q(r))} \right]^2,$$

where $\hat{\beta}_Q(r)$ is the least squares estimate of β_Q for fixed r and $\hat{\sigma}(\hat{\beta}_Q(r))$ is the estimated standard deviation of $\hat{\beta}_Q(r)$. Though for each fixed r, $t(r)$ follows a standard normal distribution asymptotically, the statistic $t(\hat{r})$ no longer follows a standard normal distribution asymptotically. The statistic $t(\hat{r})$ is essentially a likelihood ratio test statistic. If the assumption of normality were made for the $Z_i(\omega)$'s, it can be shown that $t(\hat{r})$ is

equivalent to the likelihood ratio statistic for testing the null hypothesis $\sigma_g^2 = 0$ against the alternative $\sigma_g^2 > 0$ (see 2.4.3). However, the classical asymptotic theory does not apply to this likelihood ratio test statistic since the model under the null hypothesis is not identifiable. In fact, the parameter r does not appear under the null hypothesis. This causes serious problems for determining the critical values of the test (Davies 1977, 1987).

The advantage of our model (2.7) is that it is always identifiable. The null hypothesis is equivalent to $\beta_1 = 0$ and $\beta_2 = 0$. Since the conditions for a classical regression model are satisfied, the estimates $(\hat{\beta}_1, \hat{\beta}_2)$ follow an asymptotic bivariate normal distribution. Let $\hat{\Sigma}$ be the estimated variance-covariance matrix of $(\hat{\beta}_1, \hat{\beta}_2)$. Then the Wald statistic W given below can be used to test the null hypothesis,

$$W = (\hat{\beta}_1, \hat{\beta}_2)\hat{\Sigma}^{-1}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}.$$

Under the null hypothesis, W follows an asymptotic χ_2^2 distribution.

The alternative hypothesis of the above Wald test is two-sided, i.e., $\beta_1 \neq 0$ or $\beta_2 \neq 0$. However, in the current context, we know that $\beta_1 > 0$ or $\beta_2 > 0$ under the alternative hypothesis. In other words, the alternative hypothesis is one-sided. In order to make the test more powerful against this one-sided alternative, we modify W as follows. In the computation of W, $\hat{\beta}_1$ or $\hat{\beta}_2$ are replaced by 0 whenever they are negative. Let $\hat{\sigma}_1$ and $\hat{\sigma}_2$ denote the estimated standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively, and let $\hat{\rho}$ be the estimated correlation coefficient between $\hat{\beta}_1$ and $\hat{\beta}_2$. The modified Wald statistic is

given below:

$$\tilde{W} = \begin{cases} 0 & \text{if } \hat{\beta}_1 \le 0 \text{ and } \hat{\beta}_2 \le 0, \\ \frac{\hat{\beta}_1^2}{(1-\hat{\rho}^2)\hat{\sigma}_1^2} & \text{if } \hat{\beta}_1 > 0 \text{ and } \hat{\beta}_2 \le 0, \\ \frac{\hat{\beta}_2^2}{(1-\hat{\rho}^2)\hat{\sigma}_2^2} & \text{if } \hat{\beta}_1 \le 0 \text{ and } \hat{\beta}_2 > 0, \\ \frac{\hat{\beta}_1^2}{(1-\hat{\rho}^2)\hat{\sigma}_1^2} - \frac{2\hat{\beta}_1\hat{\beta}_2\hat{\rho}}{(1-\hat{\rho}^2)\hat{\sigma}_1\hat{\sigma}_2} + \frac{\hat{\beta}_2^2}{(1-\hat{\rho}^2)\hat{\sigma}_2^2} & \text{if } \hat{\beta}_1 > 0 \text{ and } \hat{\beta}_2 > 0. \end{cases}$$

The asymptotic distribution of the modified Wald statistic $\tilde{W}$ is not standard. But it can be obtained from the asymptotic distribution of $(\hat{\beta}_1, \hat{\beta}_2)$. Let Z_1 and Z_2 be two correlated standard normal variables with correlation coefficients ρ estimated by $\hat{\rho}$. Then $\tilde{W}$ can be expressed asymptotically as

$$\tilde{W} \stackrel{D}{=} \begin{cases} 0 & \text{if } Z_1 \le 0 \text{ and } Z_2 \le 0, \\ \frac{Z_1^2}{1-\hat{\rho}^2} & \text{if } Z_1 > 0 \text{ and } Z_2 \le 0, \\ \frac{Z_2^2}{1-\hat{\rho}^2} & \text{if } Z_1 \le 0 \text{ and } Z_2 > 0, \\ \frac{Z_1^2}{1-\hat{\rho}^2} - \frac{2Z_1Z_2\hat{\rho}}{1-\hat{\rho}^2} + \frac{Z_2^2}{1-\hat{\rho}^2} & \text{if } Z_1 > 0 \text{ and } Z_2 > 0. \end{cases}$$

Let $\tilde{w}$ be the observed value of $\tilde{W}$, the p-value of the modified Wald test is computed as

$$\begin{aligned} P(\tilde{W} \ge \tilde{w}) &= 2P(Z_1 \ge \sqrt{\tilde{w}(1-\hat{\rho}^2)}, Z_2 \le 0) \\ &\quad + P(Z_1^2 - 2\hat{\rho}Z_1Z_2 + Z_2^2 \ge \tilde{w}(1-\hat{\rho}^2), Z_1 > 0, Z_2 > 0). \end{aligned}$$

The factor 2 in the first term on the right results from the symmetry of the asymptotic expression of $\tilde{W}$ in region ($Z_1 > 0$, $Z_2 \le 0$) and ($Z_1 \le 0$, $Z_2 > 0$). The probabilities on the right hand side of the above equation can be computed with the joint distribution of Z_1 and Z_2 by numerical quadrature. The formula can also be used to compute the critical values. For the $\hat{\rho}$ values ranging from -0.8 to 0.8 with an equal space 0.02, the critical values of the modified Wald test at level 0.05 are given in Table 2.4.

Table 2.4: Critical values of the modified Wald test at level $\alpha = 0.05$

ρ	c_α	ρ	c_α	ρ	c_α
-0.80	10.709	-0.26	4.575	0.28	4.131
-0.78	9.864	-0.24	4.533	0.30	4.133
-0.76	9.169	-0.22	4.497	0.32	4.133
-0.74	8.585	-0.20	4.463	0.34	4.133
-0.72	8.095	-0.18	4.428	0.36	4.138
-0.70	7.672	-0.16	4.396	0.38	4.137
-0.68	7.309	-0.14	4.369	0.40	4.143
-0.66	6.992	-0.12	4.345	0.42	4.142
-0.64	6.715	-0.10	4.320	0.44	4.147
-0.62	6.472	-0.08	4.301	0.46	4.150
-0.60	6.254	-0.06	4.281	0.48	4.160
-0.58	6.059	-0.04	4.263	0.50	4.164
-0.56	5.889	-0.02	4.247	0.52	4.168
-0.54	5.732	0.00	4.234	0.54	4.178
-0.52	5.591	0.02	4.223	0.56	4.187
-0.50	5.468	0.04	4.209	0.58	4.188
-0.48	5.353	0.06	4.197	0.60	4.202
-0.46	5.244	0.08	4.184	0.62	4.210
-0.44	5.151	0.10	4.175	0.64	4.223
-0.42	5.062	0.12	4.166	0.66	4.231
-0.40	4.982	0.14	4.160	0.68	4.237
-0.38	4.907	0.16	4.155	0.70	4.240
-0.36	4.844	0.18	4.150	0.72	4.255
-0.34	4.779	0.20	4.143	0.74	4.269
-0.32	4.722	0.22	4.142	0.76	4.285
-0.30	4.667	0.24	4.137	0.78	4.290
-0.28	4.623	0.26	4.135	0.80	4.305

2.3.5 A comparison between the modified Wald test and the ideal *t* test

In this subsection, we present the results of a simulation study. The simulation study is designed to verify the validity of the asymptotic distribution of the modified Wald statistic and to evaluate its power. The interval mapping approach can be applied using

model (2.7) with any combined response variable $Z_i(\omega)$ and it can be used for the single interval mapping as well as for the multi-point interval mapping. For the sake of convenience, without loss of generality, we only consider the setting of Fulker and Cardon (1994) for the single interval mapping in the simulation study. The modified Wald test is compared with the nominal t test considered by Fulker and Cardon. Since the β coefficient of the regression model used by Fulker and Cardon is non-positive, we restrict the estimate of β to be non-positive in minimizing the residual sum of squares, and the lower tail critical values of the standard normal distribution (-1.6449 for $\alpha = 0.05$, -2.3263 for $\alpha = 0.01$ and -3.0902 for $\alpha = 0.001$) are used.

We evaluate the type I error probability of the tests first. The type I error probabilities of the tests are evaluated for intervals with different lengths and different marker allele numbers. The interval lengths considered are 10, 20, 30, 40 and 50 cM. The marker allele numbers considered are 2, 6 and 10. The allele frequencies are taken equal. For each interval, the two flanking markers are taken to have the same number of alleles. The data are simulated under the assumption of no QTL existing in the interval. Then, under each setting, 1,000 sib pair data including the trait values of the sibs and the genotypes at the flanking markers of the sibs and their parents are simulated. The modified Wald test is carried out using its asymptotic critical values at levels 0.001, 0.01 and 0.05. This procedure is replicated 100,000 times. The proportion of rejections among these $100,000$ simulated tests is taken as the approximation to the true level of the modified Wald test. The simulated true levels of the modified Wald test are reported in the left panel of Table 2.5. The simulated true levels match perfectly the nominal

Table 2.5: Simulated actual levels of the modified Wald test and the nominal t test

No. of alleles	Interval length	Modified Wald test			Nominal t test		
		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$
2	10	0.0010	0.010	0.050	0.0030	0.025	0.105
	20	0.0010	0.010	0.051	0.0030	0.026	0.109
	30	0.0010	0.010	0.051	0.0029	0.026	0.107
	40	0.0011	0.010	0.049	0.0024	0.022	0.102
	50	0.0011	0.010	0.049	0.0024	0.021	0.092
6	10	0.0010	0.010	0.051	0.0026	0.021	0.094
	20	0.0011	0.010	0.051	0.0028	0.023	0.010
	30	0.0009	0.010	0.051	0.0025	0.024	0.103
	40	0.0011	0.010	0.051	0.0025	0.022	0.101
	50	0.0009	0.010	0.050	0.0021	0.020	0.094
10	10	0.0010	0.010	0.050	0.0021	0.020	0.089
	20	0.0009	0.010	0.050	0.0025	0.022	0.099
	30	0.0010	0.010	0.050	0.0025	0.023	0.102
	40	0.0012	0.010	0.050	0.0027	0.022	0.101
	50	0.0012	0.011	0.050	0.0024	0.020	0.096

levels. However, this is not the case for the nominal t test considered by Fulker and Carden. The true levels of the nominal t test using normal critical values are also simulated and reported in the right panel of Table 2.5. There are remarkable discrepancies between the true levels of the nominal t test and the nominal levels. The true levels of the nominal t test are more than twice the nominal levels in most of the cases. This confirms our earlier statement that the nominal t test with normal critical values is not valid in the sense that the type I error rate is not controlled as the nominal levels intend.

To evaluate the power, we compare the modified Wald test with an *ideal* likelihood ratio test. The *ideal* likelihood ratio test rejects the null hypothesis that $\sigma_g^2 = 0$ at level α if $t(\hat{r})$ is greater than its true α-level critical value. We refer to this test as the ideal t test

from now onwards. The true critical values of $t(\hat{r})$ can be simulated under the simulation settings (it must be noted that the true critical values can be simulated only under the simulation settings, they cannot be simulated in practical situations). On top of the settings for the evaluation of the type I error probability, a QTL with the same number of alleles as the flanking markers locating at the middle of each interval is simulated. Throughout the simulation, the genetic variance σ_g^2 is taken to be 0.125. This genetic variance is achieved by taking the allele contributions to the trait value as $\frac{1}{4}$ and $-\frac{1}{4}$ in the 2-allele case, $\frac{1}{4}$, $\frac{\sqrt{7}}{8}$, $\frac{1}{8}$, $-\frac{1}{4}$, $-\frac{\sqrt{7}}{8}$ and $-\frac{1}{8}$ in the 6-allele case, and $\frac{1}{4}$, $\frac{1}{8}$, $\frac{\sqrt{2}}{4}$, $\frac{\sqrt{2}}{8}$, $\frac{\sqrt{5}}{8}$, $-\frac{1}{4}$,$-\frac{1}{8}$, $-\frac{\sqrt{2}}{4}$, $-\frac{\sqrt{2}}{8}$ and $-\frac{\sqrt{5}}{8}$ in the 10-allele case. The heritability h of the QTL is set at 0.5 or 0.25, which is achieved by manipulating the value of the error variance. Both the modified Wald test and the ideal t test are simulated. Under each setting, the tests are performed with $1,000$ simulated sib pairs, and each test is repeated $100,000$ times. The simulated powers — the proportions of rejections — are reported in Table 2.6.

Overall, the modified Wald test and the ideal t test are comparable in terms of power. For shorter intervals, the ideal t test is slightly more powerful than the modified Wald test. But for longer intervals, the modified Wald test is slightly more powerful than the ideal t test. As far as we can see from the simulation results, the modified Wald test is as good as the ideal t test.

From the simulation study, we can draw the following conclusions. The modified Wald test can effectively control the type I error rate while the type I error rate is not appropriately controlled by the nominal t test considered by Fulker and Candon. The

power of the modified Wald test is comparable with the ideal t test. However, it must be emphasized that the ideal t test is only possible in the simulation and cannot be implemented in practical situations.

2.4 Technical proofs

2.4.1 Equivalence of the coefficients in $E(\pi_B \mid \pi_A, \pi_C)$ derived from the joint distribution of the IBD proportions at 3 loci and those derived by Fulker and Cardon (1994)

In section 2.2.2, we have proved that

$$\begin{aligned} E(\pi_B \mid \pi_A, \pi_C) &= \frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}} + \pi_A\left[\frac{\Psi_{AB}(1-\Psi_{BC})}{1-\Psi_{AC}} - \frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}}\right] \\ &\quad +\pi_C\left[\frac{(1-\Psi_{AB})\Psi_{BC}}{1-\Psi_{AC}} - \frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}}\right]. \end{aligned}$$

Here we are going to prove the following equalities:

$$\begin{aligned} \beta_{M1} &= \frac{\Psi_{AB}(1-\Psi_{BC})}{1-\Psi_{AC}} - \frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}} = \frac{(1-2\theta_{AB})^2 - (1-2\theta_{BC})^2(1-2\theta_{AC})^2}{1-(1-2\theta_{AC})^4}, \\ \beta_{M2} &= \frac{(1-\Psi_{AB})\Psi_{BC}}{1-\Psi_{AC}} - \frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}} = \frac{(1-2\theta_{BC})^2 - (1-2\theta_{AB})^2(1-2\theta_{AC})^2}{1-(1-2\theta_{AC})^4}, \\ \alpha_M &= \frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}} = \frac{1}{2}(1-\beta_{M1}-\beta_{M2}). \end{aligned}$$

Table 2.6: Simulated powers of the modified Wald test and the ideal t test

h	No. of alleles	Interval length	Level of the tests					
			$\alpha = 0.001$		$\alpha = 0.01$		$\alpha = 0.05$	
			t	Wald	t	Wald	t	Wald
0.5	2	10	0.780	0.765	0.940	0.933	0.980	0.970
		20	0.565	0.548	0.838	0.825	0.970	0.958
		30	0.365	0.348	0.668	0.663	0.880	0.875
		40	0.175	0.195	0.455	0.483	0.680	0.688
		50	0.043	0.088	0.203	0.325	0.445	0.575
	6	10	0.990	0.955	1.000	0.995	1.000	1.000
		20	0.925	0.900	0.995	0.990	1.000	0.998
		30	0.785	0.763	0.958	0.930	0.990	0.990
		40	0.448	0.493	0.720	0.745	0.923	0.923
		50	0.198	0.335	0.483	0.633	0.790	0.870
	10	10	0.995	0.945	1.000	1.000	1.000	1.000
		20	0.968	0.925	0.995	0.990	0.998	0.998
		30	0.838	0.780	0.953	0.940	0.998	0.993
		40	0.533	0.623	0.813	0.848	0.948	0.955
		50	0.218	0.390	0.528	0.685	0.780	0.865
0.25	2	10	0.083	0.078	0.285	0.260	0.525	0.478
		20	0.048	0.043	0.178	0.158	0.398	0.380
		30	0.020	0.025	0.133	0.120	0.333	0.318
		40	0.018	0.018	0.088	0.098	0.278	0.283
		50	0.003	0.005	0.050	0.070	0.195	0.213
	6	10	0.235	0.105	0.558	0.368	0.788	0.670
		20	0.135	0.103	0.333	0.280	0.638	0.563
		30	0.095	0.085	0.290	0.260	0.560	0.525
		40	0.048	0.055	0.200	0.193	0.423	0.418
		50	0.018	0.023	0.090	0.123	0.245	0.283
	10	10	0.273	0.098	0.590	0.340	0.830	0.665
		20	0.205	0.108	0.435	0.355	0.715	0.663
		30	0.115	0.088	0.350	0.308	0.583	0.550
		40	0.068	0.073	0.238	0.238	0.490	0.483
		50	0.013	0.025	0.063	0.110	0.278	0.318

To prove the above 3 equalities, the following equations are needed:

$$\begin{aligned} 2\Psi - 1 &= \Psi - (1-\Psi) = (1-\theta)^2 + \theta^2 - 2\theta(1-\theta) \\ &= (1-2\theta)^2, \\ 1 - 2\theta_{AC} &= (1-2\theta_{AB})(1-2\theta_{BC}), \\ \Psi_{AC} &= \Psi_{AB}\Psi_{BC} + (1-\Psi_{AB})(1-\Psi_{BC}), \\ 1 - \Psi_{AC} &= (1-\Psi_{AB})\Psi_{BC} + \Psi_{AB}(1-\Psi_{BC}). \end{aligned}$$

For the first equality,

$$\begin{aligned} \beta_{M1} &= \Psi_{AB}(1-\Psi_{BC})/(1-\Psi_{AC}) - (1-\Psi_{AB})(1-\Psi_{BC})/\Psi_{AC} \\ &= [\Psi_{AC}(1-\Psi_{BC}) - (1-\Psi_{AB})(1-\Psi_{BC})]/[\Psi_{AC}(1-\Psi_{AC})] \\ &= [\Psi_{AC}(1-\Psi_{BC}) + \Psi_{AB}\Psi_{BC} - \Psi_{AC}]/[\Psi_{AC}(1-\Psi_{AC})] \\ &= [\Psi_{AB}\Psi_{BC} - \Psi_{AC}\Psi_{BC}]/[\Psi_{AC}(1-\Psi_{AC})]. \end{aligned}$$

Furthermore,

$$\Psi_{AC}(1-\Psi_{AC}) = \frac{1}{4}[1-(1-2\Psi_{AC})^2] = \frac{1}{4}[1-(1-2\theta_{AC})^4],$$

$$\begin{aligned} &\Psi_{AB}\Psi_{BC} - \Psi_{AC}\Psi_{BC} \\ &= [\theta_{AB}^2 + (1-\theta_{AB})^2][\theta_{BC}^2 + (1-\theta_{BC})^2] - [\theta_{AC}^2 + (1-\theta_{AC})^2][\theta_{BC}^2 + (1-\theta_{BC})^2] \\ &= [\frac{1}{2} + \frac{1}{2}(1-2\theta_{AB})^2][\frac{1}{2} + \frac{1}{2}(1-2\theta_{BC})^2] - [\frac{1}{2} + \frac{1}{2}(1-2\theta_{AC})^2][\frac{1}{2} + \frac{1}{2}(1-2\theta_{BC})^2] \\ &= \frac{1}{4}(1-2\theta_{AB})^2 + \frac{1}{4}(1-2\theta_{AB})^2(1-2\theta_{BC})^2 - \frac{1}{4}(1-2\theta_{AC})^2 - \frac{1}{4}(1-2\theta_{BC})^2(1-2\theta_{AC})^2 \\ &= \frac{1}{4}[(1-2\theta_{AB})^2 - (1-2\theta_{BC})^2(1-2\theta_{AC})^2], \end{aligned}$$

thus the first equality is proved.

The second equality can be obtained easily by interchanging Ψ_{AB} and Ψ_{BC}, θ_{AB} and θ_{BC}.

The third equality can be justified by observing

$$\begin{aligned}\beta_{M1}+\beta_{M2} &= \frac{\Psi_{AB}(1-\Psi_{BC})+(1-\Psi_{AB})\Psi_{BC}}{1-\Psi_{AC}}-2\frac{(1-\Psi_{AB})(1-\Psi_{BC})}{\Psi_{AC}}\\ &= 1-2\alpha_M.\end{aligned}$$

2.4.2 Unified regression model

We have

$$\begin{aligned}\mathrm{Var}(X_{li}|\pi_{qi}) &= \mathrm{Var}(X_{li})=\sigma_g^2+\sigma_e^2, \quad l=1,2,\\ \mathrm{Cov}(X_{1i},X_{2i}|\pi_{qi}) &= \mathrm{Cov}(g_{1i},g_{2i}|\pi_{qi})+\mathrm{Cov}(e_{1i},e_{2i})\\ &= \sigma_g^2\pi_{qi}+\rho\,\sigma_e^2,\end{aligned}$$

assuming $\sigma_d^2=0$ (Lange 2002, Chapter 6). Since

$$\begin{aligned}E(Y_i^D|\pi_{qi}) &= \mathrm{Var}(X_{1i}|\pi_{qi})+\mathrm{Var}(X_{2i}|\pi_{qi})-2\mathrm{Cov}(X_{1i},X_{2i}|\pi_{qi})\\ &= 2(\sigma_g^2+\sigma_e^2-\rho\sigma_e^2)-2\sigma_g^2\pi_q,\\ E(Y_i^S|\pi_{qi}) &= \mathrm{Var}(X_{1i}|\pi_{qi})+\mathrm{Var}(X_{2i}|\pi_{qi})+2\mathrm{Cov}(X_{1i},X_{2i}|\pi_{qi})\\ &= 2(\sigma_g^2+\sigma_e^2+\rho\sigma_e^2)+2\sigma_g^2\pi_q,\end{aligned}$$

it is easy to see that

$$\begin{aligned} E(Z_i(\omega)|\pi_{qi}) &= 2(\sigma_g^2 + \sigma_e^2)(2\omega - 1) + 2\rho\,\sigma_e^2 + 2\sigma_g^2\pi_{qi} \\ &= \alpha_Q(\omega) + \beta_Q\pi_{qi}. \end{aligned}$$

where $\alpha_Q(\omega) = 2(\sigma_g^2 + \sigma_e^2)(2\omega - 1) + 2\rho\,\sigma_e^2$ and $\beta_Q = 2\sigma_g^2$.

2.4.3 Equivalence of $t(\hat{r})$ and the likelihood ratio statistic

For the linear regression model:

$$y = \alpha + \beta x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma_e^2),$$

denote the regression sum of squares with SS_R, the total sum of squares with SS_T, and the residual sum of squares with SS_E. Let

$$l_{xx} = \Sigma_i(x_i - \bar{x})^2,$$

we have:

$$\begin{aligned} SS_R &= \Sigma_i(\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 l_{xx}, \\ SS_E &= SS_T - SS_R, \\ \hat{\sigma}^2(\hat{\beta}) &= \frac{\hat{\sigma}_e^2}{l_{xx}}, \quad \text{where } \hat{\sigma_e^2} = \frac{SS_E}{n-2}. \end{aligned}$$

Thus,

$$\begin{aligned} t^2 &= \left(\frac{\hat{\beta}}{\sigma(\hat{\beta})}\right)^2 = \frac{\hat{\beta}^2}{SS_E/[(n-2)l_{xx}]} \\ &= (n-2)\frac{SS_R}{SS_E} = (n-2)(\frac{SS_T}{SS_E} - 1). \end{aligned}$$

In section 2.3.4, since SS_T is fixed with respect to the putative QTL location r and $\hat{r}$ is the value of r that minimizes SS_E, thus $\hat{r}$ is the value of r at which the maximum of $t(r)^2$ is obtained:

$$[t(\hat{r})]^2 = \max_{0 \le r \le \gamma} [t(r)]^2 = \max_{0 \le r \le \gamma} \left[\frac{\hat{\beta}_Q(r)}{\hat{\sigma}(\hat{\beta}_Q(r))} \right]^2 .$$

To test the null hypothesis $H_0 : \beta = 0$ against the alternative hypothesis $H_1 : \beta > 0$, the likelihood ratio test statistic is equivalent to $\hat{\sigma}^2_{e0}/\hat{\sigma}^2_{e1}$, where $\hat{\sigma}^2_{e0} = SS_T/n$ is the MLE of σ^2_e under H_0 and it is a constant, and $\hat{\sigma}^2_{e1} = SS_E(\hat{r})/n$ is the MLE of σ^2_e under H_1. Therefore the statistic $t(\hat{r})$ is essentially a likelihood ratio test statistic.

Chapter 3

Genome Search with Interval Mapping and the Overall Threshold

3.1 Introduction

The interval mapping methods with 2 flanking markers are more powerful in detecting the QTL than those single-marker QTL mapping methods (Lander and Botstein 1989, Haley and Knott 1992, Zeng 1994). However, in real QTL mapping, one's attention is not confined to a single interval. The searching of QTL can be genome wide, and thus multiple hypothesis testing is implicit (Haley *et al*. 1994, Jansen 1992, 1993, Jansen and Stam 1994, Zeng 1993, 1994). A problem of the multiple hypothesis testing is the difficulty of determining appropriate thresholds, and the sources of this difficulty are twofold. The first source is the problem of determining or approximating the distribu-

tion of the test statistic under the null hypothesis. The second source is that, some or all of these tests are not independent and the dependence structure of these tests is difficult to analyze (Churchill and Doerge 1994). In the genome search for experimental species, this issue has been tackled by several authors (Feingold *et al*. 1993, Rebai *et al*. 1994, 1995, Piepho 2001, Churchill and Doerge 1994, Zou *et al*. 2004, Chen and Chen 2005).

Rebai *et al*. (1994, 1995) provided an explicit formula for the upper bound of the thresholds for the backcross and F2 populations and derived a conservative threshold for single interval mapping and the searching with many intervals using the results of Davies (1977, 1987). Piepho (2001) also used the results of Davies (1977, 1987) and provided a quick method for computing the approximate thresholds for interval mapping and CIM that control the genome-wide type I error probability.

Churchill and Doerge (1994) proposed an empirical method for determining the thresholds based on the permutation test. See also Doerge and Churchill (1996). They suggested to simulate the situations under the null hypothesis by shuffling the trait values and thus destroying the potent association between the trait values and the marker genotypes, and the shuffled data are analyzed and the resulting test statistics are calculated for every analysis point (a marker locus or a marker interval, depending on the problem being investigated). The shuffling is repeated many times, and at the end of the procedure the empirical overall $100(1-\alpha)\%$ threshold that is valid simultaneously for all analysis points is estimated. Churchill and Doerge avoided the difficulty of analyzing the dependence structure of multiple hypothesis testing. Their permutation test

method can be used in both one-marker mapping and interval mapping.

Zou *et al*. (2004) proposed a resampling method to assess the genome-wide significance level of QTL mapping. They proved the score statistic can be approximated by a statistic which is a function of certain normal random variables. The threshold is then estimated by the empirical critical value through the resampling of normal random variables. They claimed this efficient resampling method is less computationally demanding than permutation tests and more accurate than theoretical approximations when rigid requirements of theoretical approximations are not satisfied.

The genome search methods briefly reviewed above are either for experimental species or based on non-interval mapping approaches. In this chapter, we deal with the genome search of QTL with interval mapping and provide an approach to determining the genome-wide thresholds.

3.2 The genome search statistic and the overall threshold

3.2.1 The genome search method with interval mapping

Multiple tests will inevitably inflate the overall type I error probability, and this can be illustrated by the following example. Suppose there are n intervals under investigation. The common significance level of the tests is set at α. We denote the modified Wald statistic for the i-th interval with W_i and its critical value c_i. The probability of claiming

the existence of QTL while in fact there is no QTL at all is:

$$\mathrm{P}(\cup_{i=1}^{n}(W_i \geq c_i)) \geq \mathrm{P}(W_1 \geq c_1) = \alpha.$$

To control the overall type I error probability, an overall threshold, c, must be used, which satisfies:

$$\mathrm{P}(\cup_{i=1}^{n}(W_i \geq c)) = \mathrm{P}(\max_{1\leq i\leq n} W_i \geq c) = \alpha,$$

where the probabilities are computed under the null hypothesis of no QTL existing.

The genome search strategy with interval mapping is as follows: first, for each interval, calculate the modified Wald statistic W_i, which was defined in Chapter 2; second, compare each W_i with the overall threshold value c, if $W_i \geq c$, then claim that a QTL exists in interval i.

Now it remains to derive the overall threshold value c, which we deal with next.

3.2.2 Calculation of the overall threshold

Suppose there are totally m intervals being investigated that are flanked by markers $\mathrm{M}_1^{(1)}$, $\mathrm{M}_2^{(1)}, \cdots, \mathrm{M}_1^{(m)}$ and $\mathrm{M}_2^{(m)}$. Some of the markers may be identical, for example, if the j-th interval and the $(j+1)$-th interval are consecutive, the right flanking marker $\mathrm{M}_2^{(j)}$ of the j-th interval and the left flanking marker $\mathrm{M}_1^{(j+1)}$ of the $(j+1)$-th interval are actually

identical. let

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}, \qquad \boldsymbol{X}_j = \begin{pmatrix} \hat{\pi}_{1,1}^{(j)} & \hat{\pi}_{2,1}^{(j)} \\ \hat{\pi}_{1,2}^{(j)} & \hat{\pi}_{2,2}^{(j)} \\ \vdots & \vdots \\ \hat{\pi}_{1,n}^{(j)} & \hat{\pi}_{2,n}^{(j)} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j,1} \\ \beta_{j,2} \end{pmatrix},$$

where Z_k is the weighted average of the squared sum and the squared difference of the trait values of the k-th sib pair, $Z_k(\omega)$, as in Chapter 2, $\hat{\pi}_{1,k}^{(j)}$ and $\hat{\pi}_{2,k}^{(j)}$ denote the proportions of alleles IBD shared by the k-th sib pair at the left and right flanking markers of the j-th interval, respectively.

For each interval, we have the model:

$$Z_k = \alpha + \hat{\pi}_{1,k}^{(j)}\beta_{j,1} + \hat{\pi}_{2,k}^{(j)}\beta_{j,2} + e, \quad e \sim \mathrm{N}(0, \sigma_e^2).$$

The $\boldsymbol{\beta}_j$ is estimated by

$$\hat{\boldsymbol{\beta}}_j = \begin{pmatrix} \hat{\beta}_{j,1} \\ \hat{\beta}_{j,2} \end{pmatrix} = (\boldsymbol{X}_j'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}')\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}')\boldsymbol{Z} = (\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_j)^{-1}\tilde{\boldsymbol{X}}_j'\boldsymbol{Z}, \tag{3.1}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{1} = (1, 1, \cdots, 1)'$ and $\tilde{\boldsymbol{X}}_j = (\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}')\boldsymbol{X}_j$.

The distribution (asymptotic distribution in the case of non-normality assumption) of $\hat{\boldsymbol{\beta}}_j$ is:

$$\hat{\boldsymbol{\beta}}_j \sim \mathrm{N}(\boldsymbol{\beta}_j, \sigma_e^2(\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_j)^{-1}).$$

Further, consider the joint distribution of $\hat{\boldsymbol{\beta}}_j$s. Let

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \cdots, \hat{\boldsymbol{\beta}}_m')'.$$

The distribution of $\hat{\boldsymbol{\beta}}$ is again a multivariate normal distribution with mean vector

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \cdots, \boldsymbol{\beta}_m')'$$

and variance-covariance matrix $\boldsymbol{\Sigma}$ given below,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}(\hat{\boldsymbol{\beta}}_1) & \mathrm{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) & \cdots & \mathrm{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_m) \\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_1) & \mathrm{Var}(\hat{\boldsymbol{\beta}}_2) & \cdots & \mathrm{Cov}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_m) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\beta}}_1) & \mathrm{Cov}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\beta}}_2) & \cdots & \mathrm{Var}(\hat{\boldsymbol{\beta}}_m) \end{pmatrix}.$$

The diagonal block entries of $\boldsymbol{\Sigma}$ are $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j) = \sigma_e^2(\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_j)^{-1}$, $j = 1, \cdots, m$. For any j and l, we have

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_l) = \sigma_e^2(\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_j)^{-1}\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_l(\tilde{\boldsymbol{X}}_l'\tilde{\boldsymbol{X}}_l)^{-1}.$$

When $l = j$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_l)$ reduces to $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j) = \sigma_e^2(\tilde{\boldsymbol{X}}_j'\tilde{\boldsymbol{X}}_j)^{-1}$.

The σ_e^2 can be obtained by averaging the estimates of σ_e^2 derived from the m regression models.

Under the null hypothesis of no QTL at all, $\hat{\boldsymbol{\beta}}$ will follow the multivariate normal distribution specified above but with mean vector $\mathbf{0}$.

Let the entries of $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j)$ be

$$\begin{pmatrix} \hat{\sigma}_{j,1}^2 & \hat{\rho}_j\hat{\sigma}_{j,1}\hat{\sigma}_{j,2} \\ \hat{\rho}_j\hat{\sigma}_{j,1}\hat{\sigma}_{j,2} & \hat{\sigma}_{j,2}^2 \end{pmatrix},$$

then the modified Wald statistic W_j is calculated using the following formula:

$$W_j = \begin{cases} 0 & \text{if } \hat{\beta}_{j,1} \le 0 \text{ and } \hat{\beta}_{j,2} \le 0, \\ \frac{\hat{\beta}_{j,1}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,1}^2} & \text{if } \hat{\beta}_{j,1} > 0 \text{ and } \hat{\beta}_{j,2} \le 0, \\ \frac{\hat{\beta}_{j,2}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,2}^2} & \text{if } \hat{\beta}_{j,1} \le 0 \text{ and } \hat{\beta}_{j,2} > 0, \\ \frac{\hat{\beta}_{j,1}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,1}^2} - \frac{2\hat{\beta}_{j,1}\hat{\beta}_{j,2}\hat{\rho}_j}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,1}\hat{\sigma}_{j,2}} + \frac{\hat{\beta}_{j,2}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,2}^2} & \text{if } \hat{\beta}_{j,1} > 0 \text{ and } \hat{\beta}_{j,2} > 0. \end{cases}$$

Since $\max_{1\le j\le n} W_j$ is a function of $\hat{\boldsymbol{\beta}}$, the overall threshold c such that

$$\mathrm{P}(\max_{1\le j\le n} W_j \ge c \,|\, \mathrm{H}_0) = \alpha$$

can be simulated by the following procedure:

1. generate a multivariate normal random variable Y with mean $(0, 0, \cdots, 0)'$ and variance-covariance matrix $\boldsymbol{\Sigma}$, $Y = (Y_{11}, Y_{12}, Y_{21}, Y_{22}, \cdots, Y_{m1}, Y_{m2})'$;

2. compute the modified Wald statistic w_j with (Y_{j1}, Y_{j2}),

$$w_j = \begin{cases} 0 & \text{if } Y_{j1} \le 0 \text{ and } Y_{j2} \le 0, \\ \frac{Y_{j1}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,1}^2} & \text{if } Y_{j1} > 0 \text{ and } Y_{j2} \le 0, \\ \frac{Y_{j2}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,2}^2} & \text{if } Y_{j1} \le 0 \text{ and } Y_{j2} > 0, \\ \frac{Y_{j1}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,1}^2} - \frac{2Y_{j1}Y_{j2}\hat{\rho}_j}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,1}\hat{\sigma}_{j,2}} + \frac{Y_{j2}^2}{(1-\hat{\rho}_j^2)\hat{\sigma}_{j,2}^2} & \text{if } Y_{j1} > 0 \text{ and } Y_{j2} > 0. \end{cases}$$

   and record $T_1 = \max_{1\le j\le m} w_j$;

3. repeat step 1 and 2 for a number of times, for example 500 times, and we have a sample $(T_1, T_2, \cdots, T_{500})$ from the null distribution of $\max_{1\le j\le m} W_j$;

4. compute the empirical $100(1-\alpha)\%$ percentile of the sample T, and this is the simulated threshold value c.

3.3 Simulation studies

We assume the random errors of the sib pair trait values follow a bivariate normal distribution with mean $(0, 0)'$ and variance-covariance matrix $\begin{pmatrix} \sigma_e^2 & \rho\,\sigma_e^2 \\ \rho\,\sigma_e^2 & \sigma_e^2 \end{pmatrix}$, where the correlation coefficient ρ is fixed at 0.3.

The power of the genome search with interval mapping is assessed for both single QTL case and multiple QTL case.

For the single QTL case, We simulated an 80cM chromosome segment with 9 markers equally spaced at 10cM on it. The 9 markers have the same number of alleles, say 3, 6 or 10, and the alleles at each marker locus are equally frequent. A single QTL locating at 32.5cM or 72.5cM from the left that has 3 equally frequent alleles is simulated on the 80cM chromosome segment. The layout of the markers and the QTL is demonstrated by figure 3.1. The environmental variance σ_ϵ^2 is set at 0.5. The genetic variance σ_g^2

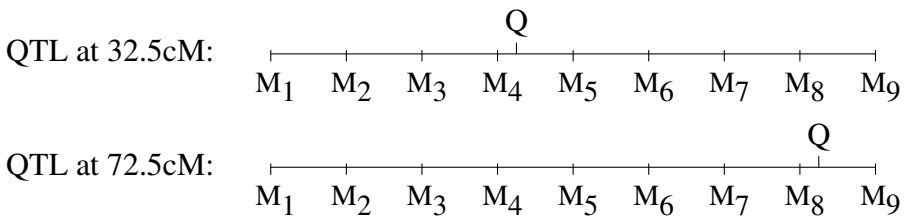


Figure 3.1: Layout of the markers and the QTL – single QTL

takes either 0.5, which is achieved by taking ($\frac{\sqrt{6}}{4}$, 0, -$\frac{\sqrt{6}}{4}$) as the allele contributions to the trait value, or 0.25, which is achieved by taking the allele contributions ($\sqrt{0.1875}$, 0, - $\sqrt{0.1875}$), and the heritability h is $\frac{1}{2}$ and $\frac{1}{3}$ accordingly. The QTL genotypes are

generated together with the marker genotypes, because they are necessary for generating the sib pair trait values. The procedure for generating the sib pair trait values and genotypes, including the markers' and the QTL's, is as follows.

1. *Generating the parental genotypes.* The generated parental genotypes are stored in a matrix A consisting of 4 columns, where the first 2 columns represent the two haplotypes of the father and the other 2 columns represent the haplotypes of the mother. The number of rows l is the number of loci. Let n_k be the allele number of the k-th locus. Encode the alleles of the k-th locus with numbers 1, 2, $\cdots$, n_k, and denote their corresponding frequencies with $p_{k,1}, p_{k,2}, \cdots, p_{k,n_k}$. A random number r_k taking value 1, 2, $\cdots$, n_k with probabilities $(p_{k,1}, p_{k,2}, \cdots, p_{k,n_k})$ can be thought of as a multinomial random variable. Generate 4 such multinomial random numbers with replacement, and they are the entries in the k-th row of the matrix. This procedure is carried out for every locus (row), and the parents' genotypes are generated.

2. *Generating the sib pair's genotypes.* The generated sib pair's genotypes are also stored in a matrix of 4 columns and l rows. Let θ_k denotes the recombination fraction between the k-th and the $(k+1)$-th loci. Denote the number of the column (haplotype) from which the father transmits an allele to sib 1 with c (c takes either 1 or 2). The procedure for simulating the haplotype of sib 1 inherited from the father is as follows. (a) For the first locus, generate a uniform $\mathcal{U}(0,1)$ random number r. If $r < 0.5$, the allele in the first haplotype of the father $(A_{1,1})$

is transmitted to sib 1, and $c = 1$. Otherwise the allele in the second haplotype of the father ($A_{1,2}$) is transmitted, and $c = 2$. (b) For the $(k + 1)$-th locus, generate a uniform $\mathcal{U}(0, 1)$ random number r. if $r < \theta_k$, c is updated with (3-c) so that the number of the column from which to transmit the allele at the $(k + 1)$-th locus is changed from 1 to 2 or from 2 to 1, otherwise, c remains the same as it was at the k-th locus. The allele in haplotype c (updated c) at the $(k+1)$-th locus of the father ($A_{k+1,c}$) is then transmitted to sib 1. (c) Repeat step (b) for $k = 1, 2, \cdots, l - 1$, and the haplotype of sib 1 inherited from the father is generated. The haplotype of sib 1 inherited from the mother and the two haplotypes of sib 2 are simulated in the same way.

3. *Generating the trait values.* First, generate the random errors of the sib pair trait values, which follow the bivariate normal distribution as stated above with $\sigma^2_\epsilon = 0.5$. Then the trait value of each sibling is generated by adding up the allele contributions corresponding to the two alleles at the QTL and its random error.

4. *Generating the sample.* Repeat the above 3 steps for 500 times.

Five hundred sib pairs are generated under each combination of QTL location (32.5cM or 72.5cM), h ($\frac{1}{2}$ or $\frac{1}{3}$) and marker allele number (3, 6, or 10). The QTL genotypes are left out of account during the genome search as they are assumed unobservable. The genome search is then conducted with marker intervals of 10cM and 20cM. When searching with 10cM intervals, each interval is determined by two adjacent markers. When searching with 20cM intervals, the intervals are determined by every other mark-

ers, say marker 1, marker 3, marker 5, *etc*.. For each set of 500 sib pair data, the variance-covariance matrix $\mathbf{\Sigma}$ of $\boldsymbol{\beta}$ is computed, and then 500 iid multivariate normal random variables, Ys, with mean $(0, 0, \cdots, 0)'$ and variance-covariance matrix $\mathbf{\Sigma}$ are generated and the statistic $T = \max_{1 \le j \le n} W_j$ is computed for each Y, and the empirical critical value of the 500 Ts is taken as the overall threshold. The generated threshold is compared to the test statistic computed with the set of 500 sib pair data. We claim that the QTL is detected if, for at least one interval, the modified Wald test is significant and the QTL location estimate in that interval is within 5cM regarding the true QTL location. Under each setting, the whole procedure is repeated 2,000 times, and the proportions that the QTL is detected are reported in Table 3.1.

Table 3.1: Simulated powers of the genome search – single QTL

QTL location	h	Allele numbers	$\alpha = 0.01$		$\alpha = 0.05$	
			10cM	20cM	10cM	20cM
32.5cM	$\frac{1}{2}$	3	0.531	0.381	0.746	0.549
		6	0.617	0.486	0.821	0.631
		10	0.613	0.488	0.825	0.619
	$\frac{1}{3}$	3	0.141	0.100	0.296	0.210
		6	0.139	0.112	0.319	0.237
		10	0.140	0.129	0.332	0.255
72.5cM	$\frac{1}{2}$	3	0.541	0.375	0.743	0.536
		6	0.596	0.460	0.810	0.604
		10	0.606	0.500	0.831	0.637
	$\frac{1}{3}$	3	0.127	0.095	0.281	0.207
		6	0.163	0.120	0.333	0.250
		10	0.146	0.124	0.331	0.259

In Table 3.1, as expected, the power decreases as the interval length increases from

10cM to 20cM. The reason is as follows. When the interval length is 10cM, interval (M_3, M_4) and (M_4, M_5) are examined if the QTL locates at 32.5cM, and interval (M_7, M_8) and (M_8, M_9) are examined if the QTL locates at 72.5cM. However, when the interval length is 20cM, only interval (M_3, M_5) is examined if the QTL locates at 32.5cM, and only interval (M_7, M_9) is examined if the QTL locates at 72.5cM. In other words, less markers are used when the genome search is conducted at 20cM intervals. It also can be seen that the position (near the middle or near the end) of the QTL in the region being examined is irrelevant to the power. Moreover, an obvious increase in power can be observed when the heritability increases from 1/3 to 1/2. It is observed that the power increases systematically when the marker allele number increases from 3 to 6, but this is not true when the marker allele number increases from 6 to 10. The reason is as follows. The proportion of π_M being uniquely determined is 37.04% for 3-allele markers, 67.13% for 6-allele markers and 80.10% for 10-allele markers (Fulker and Cardon 1994). The increase in this proportion is greater when the marker allele number increases from 3 to 6 than from 6 to 10. Therefore, when the marker allele number increases from 3 to 6, the increase in the power of the genome search is large enough and the increasing trend is not affected by random fluctuation. However, when the marker allele number increases from 6 to 10, the increase in the power is so small that the increasing trend is ruined by random fluctuation.

Next, we consider the case when 2 QTLs present simultaneously in the 80cM chromosome segment. The situations for the markers are the same as above except that only 3-allele markers are considered. The environmental variance σ_ϵ^2 is still set at 0.5. The

total genetic variance is set at 0.5, and thus the heritability is fixed at 0.5. The locations of the 2 QTLs (Q_1, Q_2) are (12.5cM, 32.5cM) or (12.5cM, 72.5cM). The layout of the markers and the QTLs are as follows. Two situations of the genetic variances

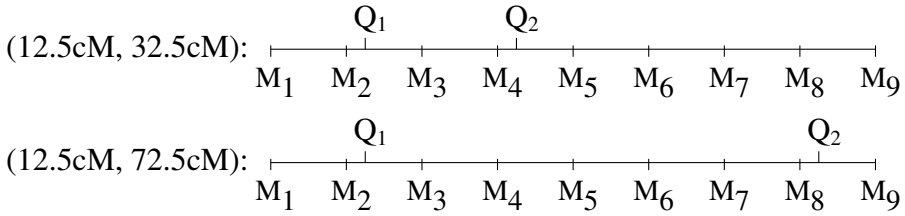


Figure 3.2: Layout of the markers and the QTLs – 2 linked QTLs

(σ^2_{g1}, σ^2_{g2}) are considered. Situation 1: the 2 QTLs have equal number of alleles, say 3 alleles, but different genetic variances, $\sigma^2_{g1} = 0.1$, achieved by taking the allele contributions ($\sqrt{0.075}$, 0, $-\sqrt{0.075}$), and $\sigma^2_{g2} = 0.4$, achieved by taking the allele contributions ($\sqrt{0.3}$, 0, $-\sqrt{0.3}$). Situation 2: the 2 QTLs have different number of alleles, 3 for Q_1 and 6 for Q_2, but equal genetic variance, $\sigma^2_{g1} = \sigma^2_{g2} = 0.25$, and the allele contributions are ($\sqrt{0.1875}$, 0, $-\sqrt{0.1875}$) for Q_1 and ($\sqrt{0.325}$, 0.2, 0.1, -0.1, -0.2, $-\sqrt{0.325}$) for Q_2. The alleles at the same locus are equally frequent. 500 sib pairs are generated under each combination of QTL location and genetic variances. The genome search is conducted with marker intervals of 10cM and 20cM. The procedures for generating the genotypes and the sib pair trait values are similar to those in the case of a single QTL. The overall thresholds are simulated in the same way as above. Under each setting, the whole procedure is replicated 2,000 times, and the proportions that the QTLs are detected at significance level 0.05 are reported in Table 3.2.

From Table 3.2, we find that when Q_2 has greater effect (σ^2_{g2}=0.4) than Q_1, its

Table 3.2: Simulated powers of the genome search – 2 linked QTLs

QTL locations	$(\sigma^2_{g1}, \sigma^2_{g2})$	Q_1 detected		Q_2 detected		both detected		either detected	
		10cM	20cM	10cM	20cM	10cM	20cM	10cM	20cM
(12.5, 32.5)	(0.1, 0.4)	0.253	0.277	0.619	0.457	0.185	0.167	0.687	0.567
	(0.25, 0.25)	0.451	0.405	0.440	0.343	0.238	0.171	0.653	0.577
(12.5, 72.5)	(0.1, 0.4)	0.043	0.039	0.540	0.365	0.026	0.017	0.557	0.387
	(0.25, 0.25)	0.211	0.155	0.212	0.172	0.045	0.028	0.378	0.299

proportions of being detected are higher than those of Q_1. By contrast, when Q_1 and Q_2 have equal effect (0.25), they have similar chances of being detected. Their proportions of being detected increase as the 2 QTLs get closer. This is because, the correlations between Z and π_M at markers close to the QTLs are increased when the 2 QTLs are more closely linked, so that the regression coefficients and then the values of the modified Wald statistics are increased. This indicates that clustered QTLs can be detected more easily. When Q_1 has a small effect (0.1) and the QTL locations are 12.5cM and 72.5cM, the proportion of Q_1 being detected is even smaller than the level of the test–0.05. However, when the QTL locations are 12.5cM and 32.5cM, the proportion of Q_1 being detected increases dramatically, from around 0.04 to around 0.26. These indicate that a small-effect QTL can hardly be detected if it is not closely linked to any large-effect QTL, and its proportion of being detected can be largely increased when it is close to a large-effect QTL. We can also observe that, whatever the QTL effects are, the proportion of both QTLs being detected is very small (< 0.05) when the QTL locations are 12.5cM and 72.5cM, but it is moderate when the QTL locations are 12.5cM and 32.5cM. This indicates that 2 QTLs can hardly be detected simultaneously if they are far apart. It

is also found that, the proportion of being detected decreases as the interval length increases, except for Q_1 when the QTL locations are (12.5cM, 32.5cM) and (σ^2_{g1}, σ^2_{g2}) is (0.1, 0.4). This exception could be explained as follows. Since σ^2_{g1} is much smaller than σ^2_{g2}, Z is mostly determined by Q_2. When Q_2 is very close to Q_1, the correlations between Z and π_M at M_3, M_2 and M_1 are also large, and the modified Wald statistic in interval (M_2, M_3) is greater than those in interval (M_1, M_2) and interval (M_1, M_3). However, the thresholds are determined by the correlations between π_M at the markers and are not affected by the relative distance of the 2 QTLs.

Finally, we consider the case of 2 unlinked QTLs. Suppose we are interested in 3 segments in different chromosomes: a 20cM segment with 3 markers equally spaced at 10cM and no QTL in Chromosome 1, a 40cM segment with 5 markers equally spaced at 10cM and one QTL (Q_1) locating at 7.5cM in Chromosome 2 and a 40cM segment with 5 markers equally spaced at 10cM and one QTL (Q_2) locating at 17.5cM in Chromosome 3. The layout is as follows. All markers have equal number of alleles, say 3 or

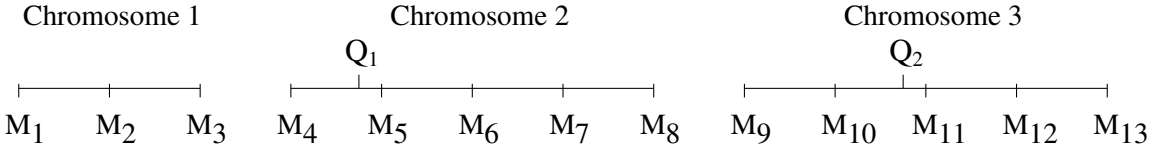


Figure 3.3: Layout of the markers and the QTLs – 2 unlinked QTLs

10, which are equally frequent. The environmental variance is set at 0.5. The total genetic variance is 0.5, and the same two situations of the genetic variances of the 2 QTLs as above are considered. The genotypes in each segment are generated separately. Five hundred sib pairs are generated under each combination of marker allele number and

genetic variance situation. The genome search is conducted with marker intervals of 10cM and 20cM. The overall thresholds are simulated in the same way as above. Under each setting, the whole procedure is repeated 2,000 times, and the proportions that the QTLs being detected are reported in Table 3.3.

Table 3.3: Simulated powers of the genome search – 2 unlinked QTLs

$(\sigma^2_{g1}, \sigma^2_{g2})$	Allele number	Q_1 detected		Q_2 detected		both detected	
		10cM	20cM	10cM	20cM	10cM	20cM
(0.1, 0.4)	3	0.024	0.014	0.444	0.440	0.012	0.006
	10	0.027	0.013	0.491	0.646	0.014	0.009
(0.25, 0.25)	3	0.127	0.083	0.132	0.155	0.015	0.011
	10	0.144	0.101	0.155	0.264	0.021	0.026

It can be found that, the proportion of being detected increases as the marker allele number increases from 3 to 10, except that of Q_1 when its genetic effect is 0.1 and the interval length is 20cM. This may arise from the combined effect of the small genetic effect of Q_1 (0.1) and the long interval (in this case, interval (M_4, M_6)), since the flanking markers of long interval contain only little QTL information and thus their polymorphism situation cannot have large impact on the QTL detection. It seems to our surprise that, for Q_2, the proportion of being detected increases largely when the interval length increases from 10cM to 20cM in 3 of the 4 cases considered here. The possible reason is that, in the 10cM case, the correlation between π_M at adjacent markers are greater and the threshold becomes much higher. It can also be observed that, when the 2 QTLs have equal effect (0.25) and the interval length is 20cM, the proportion of Q_2 being detected

is much higher than that of Q_1. This is because, only 1 interval–(M_4, M_6)–is examined for detecting Q_1, while for detecting Q_2, 2 intervals –(M_{10}, M_{11}) and (M_{10}, M_{11})–are examined and more marker information is used.

From the simulation studies, we can see that, the probability of QTL being detected is affected simultaneously by the interval length, the genetic variance, and the relative distance between QTLs if there are more than 1 QTL. For short intervals(<10cM), the probability of QTL being detected may not increase as interval length decreases, because the thresholds may increase sharply as the correlation between π_Ms increases.

Chapter 4

Multi-point Interval Mapping

In real QTL mapping problems, to confirm whether a QTL associating with a QT exists and to find out its location if it does exist, a common practice is to genotype dozens of or even more markers and conduct interval mapping interval by interval. When the two flanking markers are completely informative, that is their IBD proportions can be uniquely determined, all of the QTL information is contained in the flanking markers. Otherwise, only a part of the QTL information is contained in the flanking markers, and the rest is contained in some nearby markers. This is well demonstrated by Table I of Fulker *et al*. (1995), in which, when the flanking markers are not completely informative, the estimated $\hat{\pi}_q$ depends also on markers beyond the two flanking ones.

The interval mapping method proposed in Chapter 2 only makes use of the two flanking markers no matter they are completely informative or not. To distinguish the interval mapping method in Chapter 2 from the one to be presented in this chapter,

we will call this method "the two-point interval mapping method" and call $\hat{\pi}_M$ at each marker estimated with its own information (the sib pair's and the parents' genotypes at that single locus) alone "the local estimate" from now onwards. When the allele numbers of both flanking markers are large enough, the flanking markers will be completely informative in most of the times, and therefore the two-point interval mapping method could perform well enough. However, when applied to intervals flanked by markers with fewer alleles, for example only 2 or 3 alleles, the information contained in the two flanking markers is so little that the two-point interval mapping method tends to waste a larger part of the information, which is carried by other markers, and could become less powerful. Due to these drawbacks of the two-point interval mapping, new QTL mapping methods which make use of the available marker information as much as possible are in order.

Many authors explored ways to extract information from multiple markers. Fulker *et al*.(1995) provided a regression based multi-point interval mapping method, where the IBD proportion at the QTL is estimated by a linear combination of the IBD proportions at multiple markers. Lander and Green(1987) proposed a hidden Markov chain approach for multi-point likelihood calculation. Kruglyak *et al*.(1995) improved Lander and Green's algorithm and implemented the new algorithm in the computer package MAPMARKER/HOMOZ.

In this chapter, we are going to propose a new interval mapping method which takes all available markers into consideration. It is an extension of the interval mapping

model 2.7 in subsection 2.3.2 of Chapter 2. Instead of estimating the IBD proportions at the flanking markers using only the information in the flanking markers, the IBD proportions at the flanking markers are estimated using information in multiple markers. We call this interval mapping method "the multi-point interval mapping method".

4.1 Interval mapping model with multiple markers

In Chapter 2, the regression model proposed for the two-point interval mapping is

$$Z_i(\omega) = \alpha + \beta_1 \hat{\pi}_{1i} + \beta_2 \hat{\pi}_{2i} + e_i, \tag{4.1}$$

where $\hat{\pi}_{1i}$ and $\hat{\pi}_{2i}$ are the local estimates of the IBD proportions at the flanking markers. In the multi-point interval mapping model, these estimates are to be replaced by the multi-point estimates that are to be discussed in the following.

Suppose n marker loci are genotyped $(M_1, M_2, \cdots, M_n)$, and we denote the information in the j-th marker of the i-th sib pair with $G_{j,i}$. If the putative QTL lies in the t-th interval, we propose to use the following regression model for the multi-point mapping in the t-th interval:

$$Z_{t,i}(\omega) = \alpha_t + \beta_{t,1} E(\pi_{t,i}|G_{1,i}, \cdots, G_{t,i}) + \beta_{t,2} E(\pi_{t+1,i}|G_{t+1,i}, \cdots, G_{n,i}) + e_i. \tag{4.2}$$

The IBD proportion at the left flanking marker is estimated using information in all markers to the left of the QTL, and similarly the IBD proportion at the right flanking marker is estimated using information in all markers to the right of the QTL. For simplicity, we call these estimates "the multi-point estimates", and call the corresponding

interval mapping "the multi-point interval mapping". Note that the multi-point estimate of π_{M_1} for the leftmost marker is identical to its local estimate. Similarly, the multi-point estimate of π_{M_n} for the rightmost marker is identical to its local estimate.

4.2 Multi-point estimate of the IBD proportion at the flanking marker

In Chapter 2, a linear combination method for estimating the IBD proportion at any location within an interval using the flanking markers' information is introduced, which was first proposed by Fulker and Cardon (1994) (see detailed description in subsection 2.3.1) and proved to be correct in section 2.2 of this thesis. This method was further extended to include the information in more nearby markers by Fulker *et al*. (1995), and can be applied here. This linear combination method will be described in detail in subsection 4.2.1.

In this thesis, we present a new multi-point interval mapping method , which uses the joint distribution of the numbers of alleles IBD at multiple markers to estimate the IBD proportions at the flanking markers and then performs the two-point interval mapping procedure as presented in Chapter 2. The joint distribution of the numbers of alleles IBD at multiple markers is derived by adding up the probabilities of all possible allele-transmission patterns conditioning on the marker genotypes. The estimated IBD proportions we obtain are specific to particular marker genotypes, and thus should be

more accurate than those obtained through the linear combination approach and the hidden Markov chain approach. The details of the multi-point interval mapping method are given in section 4.2.2.

4.2.1 Estimation by linear combination

Fulker *et al*.(1995) proposed a linear combination estimate for the IBD proportion at any putative QTL location in the interval spanned by multiple markers using the information in all available markers on the same chromosome. Their procedure is as follows.

For a putative QTL located at position c and a group of markers at positions L_1, L_2, $\cdots$, L_p on the same chromosome, the estimate of π_c can be expressed by the following linear combination:

$$\hat{\pi}_c = b_0 + b_1\hat{\pi}_{L_1} + b_2\hat{\pi}_{L_2} + \cdots + b_p\hat{\pi}_{L_p},$$

where the $\hat{\pi}_{L_j}$s on the right hand side of the equation are the local estimates. By calculating the covariance between $\hat{\pi}_c$ and $\hat{\pi}_{L_j}$, the following normal equations hold:

$$\begin{pmatrix} \mathrm{Cov}(\hat{\pi}_{L_1},\hat{\pi}_c) \\ \mathrm{Cov}(\hat{\pi}_{L_2},\hat{\pi}_c) \\ \vdots \\ \mathrm{Cov}(\hat{\pi}_{L_p},\hat{\pi}_c) \end{pmatrix} = \begin{pmatrix} \mathrm{V}(\hat{\pi}_{L_1}) & \mathrm{Cov}(\hat{\pi}_{L_1},\hat{\pi}_{L_2}) & \cdots & \mathrm{Cov}(\hat{\pi}_{L_1},\hat{\pi}_{L_p}) \\ \mathrm{Cov}(\hat{\pi}_{L_2},\hat{\pi}_{L_1}) & \mathrm{V}(\hat{\pi}_{L_2}) & \cdots & \mathrm{Cov}(\hat{\pi}_{L_2},\hat{\pi}_{L_p}) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(\hat{\pi}_{L_p},\hat{\pi}_{L_1}) & \mathrm{Cov}(\hat{\pi}_{L_p},\hat{\pi}_{L_2}) & \cdots & \mathrm{V}(\hat{\pi}_{L_p}) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}.$$

For simplicity, we denote the above normal equations with $\mathbf{C} = \mathbf{VB}$.

The off-diagonal elements of **V** have the expectation

$$E[\mathrm{Cov}(\hat{\pi}_{L_i}, \hat{\pi}_{L_j})] = 8\mathrm{V}(\hat{\pi}_{L_i})\mathrm{V}(\hat{\pi}_{L_j})(1-2\theta_{ij})^2,$$

and the elements of **C** have the expectation

$$E[\mathrm{Cov}(\hat{\pi}_{L_i}, \hat{\pi}_c)] = \mathrm{V}(\hat{\pi}_{L_i})(1-2\theta_i)^2,$$

where θ_{ij} is the recombination fraction between loci L_i and L_j, and θ_i is the recombination fraction between loci L_i and c. These expectations follow from simple rearrangement of the formulas for the correlation between estimated proportions of alleles IBD shared by the sib pair, which were first derived by Elston and Keats (1985). The $V(\hat{\pi}_{L_i})$s are estimated by the empirical variances given by $\hat{V}(\hat{\pi}_{L_i}) = \frac{1}{n-1}\sum_{j=1}^{n}(\hat{\pi}_{L_i,j} - \bar{\hat{\pi}}_{L_i})^2$, where $\hat{\pi}_{L_i,j}$ is the local estimate of the IBD proportion at marker L_i of individual j .

The coefficients $b_1, b_2, \cdots, b_p$ can then be estimated by solving

$$E(\mathbf{C}) = E(\mathbf{V})\mathbf{B},$$

and b_0 is estimated as

$$\hat{b}_0 = \bar{\hat{\pi}}_c - \hat{b}_1\bar{\hat{\pi}}_1 - \hat{b}_2\bar{\hat{\pi}}_2 - \cdots - \hat{b}_p\bar{\hat{\pi}}_p.$$

The value of $\bar{\hat{\pi}}_c$ has to take the theoretical value 0.5 since the locus c is assumed to be the putative QTL and we have no information about the genotypes at the QTL. The $\bar{\hat{\pi}}_i$ can take the empirical mean of $\hat{\pi}_i$ at locus L_i.

For the multi-point interval mapping model (4.2), to estimate the IBD proportion at the left flanking marker M_t, we just need to let the locus c be M_t and calculate the multi-point estimate of π_t with the local estimates $\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_t$, to estimate the proportion of

IBD shared alleles at the right flanking marker M_{t+1}, let the locus c be M_{t+1} and calculate the multi-point estimate of π_{t+1} with the local estimates $\hat{\pi}_{t+1}, \hat{\pi}_{t+2}, \cdots, \hat{\pi}_n$.

4.2.2 Estimation by the joint density of the IBD proportions at multiple markers

In this section, we are going to propose a new method for estimating the IBD proportion at any marker using the joint density of the numbers of alleles IBD at multiple markers conditioning on the observed marker genotypes. The joint density of the numbers of alleles IBD at multiple markers introduced here is an extension to that introduced in Chapter 2, which considers only 3 loci.

Given the parents' and the sib pair's genotypes, the possible transmission patterns that imply how the parents transmit alleles to the sib pair can be determined and the corresponding probabilities can be calculated, and then the possible numbers of alleles IBD at each marker and their joint conditional distribution can be derived. Once the joint conditional distribution given the sib pair's and their parents' marker genotypes is obtained, the proportion of alleles IBD shared by the sib pair at a particular marker can be estimated by the conditional marginal expectation at that marker. In the following, we explain in detail how this estimate is calculated.

1. Let G_p be the genotypes of the parents. Let l be the number of heterozygous loci of one parent, then given the genotype of the parent, there are 2^{l-1} possible phase

known genotypes, each with equal probability $\frac{1}{2^{l-1}}$.

For example, 4 marker loci A, B, C and D are genotyped, and the parents' genotypes are:

$$\text{father}: \quad a_1a_4/b_2b_2/c_3c_4/d_1d_3, \qquad \text{mother}: \quad a_2a_3/b_1b_3/c_1c_1/d_2d_4.$$

The father has 4 possible phase known genotypes and the mother has 4 possible phase known genotypes.

2. Let G_s be the the genotypes of the sib pair. Given both parents' phase known genotypes, the genotype of one child may arise from different patterns of allele transmission.

   For the above example, one of the parents' phase known genotypes (there are 16 of them) are:

father					mother				
a_1	b_2	c_4	d_3	[1]	a_2	b_3	c_1	d_2	[1]
a_4	b_2	c_3	d_1	[2]	a_3	b_1	c_1	d_4	[2]

   where each column corresponds to one locus, the letters with a numeric subscript (*e.g.* a_2, b_3) denote the alleles at different loci, and the alleles above the line are on one chromosome (chromosome '[1]'), and the alleles below the line are on the second chromosome (chromosome '[2]'). Suppose the sib pair's genotypes are:

$$\text{sib1}: \quad a_1a_2/b_2b_3/c_1c_3/d_3d_4, \qquad \text{sib2}: \quad a_1a_2/b_1b_2/c_1c_3/d_2d_3.$$

   The possible patterns of allele transmission resulting in the given genotypes of sib1 and sib2 are given in the following table:

Table 4.1: Allele transmission patterns of the sib pair given the parents' phase known genotypes

sib1					sib2				
1	1	2	1	(F)	1	1	2	1	(F)
1	1	1	2	(M)	1	2	1	1	(M)
1	2	2	1	(F)	1	2	2	1	(F)
1	1	1	2	(M)	1	2	1	1	(M)
1	1	2	1	(F)	1	1	2	1	(F)
1	1	2	2	(M)	1	2	2	1	(M)
1	2	2	1	(F)	1	2	2	1	(F)
1	1	2	2	(M)	1	2	2	1	(M)

where the numbers above the line denote the transmission patterns of the father's alleles, and the numbers below the line denote the transmission patterns of the mother's genotype. The 1s and 2s indicate from which chromosome the alleles at each locus are transmitted to the child by the father or the mother. Thus for the first transmission pattern of sib1, the father transmits the allele on the first chromosome at locus A(a_1), the allele on the first chromosome at locus B(b_2), the allele on the second chromosome at locus C(c_3), and the allele on the first chromosome at locus D(d_3) to sib1, the mother transmits the allele on the first chromosome at locus A(a_2), the allele on the first chromosome at locus B(b_3), the allele on the first chromosome at locus C(c_1), and the allele on the second chromosome at locus D(d_4) to sib1, so that sib1 has the genotype $a_1a_2/b_2b_3/c_1c_3/d_3d_4$. The frequency of each transmission pattern of one child is the product of the frequencies of the transmission patterns from both parents to that child.

3. Let the allele transmission pattern of the sib pair at locus t be: j_1/j_2 and k_1/k_2, where j_1 and k_1 are origins of the alleles at locus t transmitted by the father, and j_2 and k_2 are origins of the alleles at locus t transmitted by the mother. j_1, j_2, k_1 and k_2 take either 1 or 2. Then the number of alleles IBD shared by the sib pair at locus t is:

$$i_t = I(j_1 = k_1) + I(j_2 = k_2). \tag{4.3}$$

4. Let hap_p be a generic notation for the phase known parents genotypes, and *tran* be a generic notation for a transmission pattern. The joint probability of $(i_1, i_2, \cdots, i_n)$, the number of alleles IBD shared at marker loci $M_1, M_2, \cdots, M_n$ by the sib pair, is given by

$$\begin{aligned} P(i_1, i_2, \cdots, i_n | G_s, G_p) &= \frac{P(i_1, i_2, \cdots, i_n, G_s, G_p)}{P(G_s, G_p)} = \frac{\sum\limits_{hap_p} P(i_1, i_2, \cdots, i_n, G_s, hap_p)}{\sum\limits_{hap_p} P(G_s, hap_p)} \\ &= \frac{\sum\limits_{hap_p} P(i_1, i_2, \cdots, i_n, G_s | hap_p)}{\sum\limits_{hap_p} P(G_s | hap_p)} = \frac{\sum\limits_{hap_p} \sum\limits_{tran_s} P(i_1, i_2, \cdots, i_n, tran_s)}{\sum\limits_{hap_p} \sum\limits_{tran_s} P(tran_s)} \\ &= \frac{\sum\limits_{hap_p} \sum\limits_{tran_s} P(i_1, i_2, \cdots, i_n | tran_s) P(tran_s)}{\sum\limits_{hap_p} \sum\limits_{tran_s} P(tran_s)}, \end{aligned}$$

where the sum $\sum hap_p$ is over all possible phase known genotypes of the parents given G_p, and the sum $\sum tran_s$ is over all possible transmission patterns given G_s and hap_p.

5. The expected number of alleles IBD shared at locus t given G_s and G_p is:

$$\begin{aligned} E(i_t|G_s,G_p) &= \sum_{i_u \neq i_t}[P(i_1,\cdots,i_{t-1},i_t=1,i_{t+1},\cdots,i_n|G_s,G_p) \\ &\quad +2P(i_1,\cdots,i_{t-1},i_t=2,i_{t+1},\cdots,i_n|G_s,G_p)] \\ &= \frac{1}{\sum_{hap_p}\sum_{tran_s} P(tran_s)} \sum_{hap_p}\sum_{tran_s} P(tran_s)\cdot X, \end{aligned}$$

where

$$\begin{aligned} X &= \sum_{i_u \neq i_t}[P(i_1,\cdots,i_{t-1},i_t=1,i_{t+1},\cdots,i_n|tran_s)+2P(i_1,\cdots,i_{t-1},i_t=2,i_{t+1},\cdots,i_n|tran_s)] \\ &= \begin{cases} 0 & if\ i_t=0 \\ 1 & if\ i_t=1 \\ 2 & if\ i_t=2 \end{cases} = i_t, \end{aligned}$$

i_t is computed by equation 4.3.

A common feature of this multi-point estimate and the local estimate of the IBD proportion at a marker locus is that, when the marker is completely informative, both estimates equal the exact IBD proportion.

We make two remarks to conclude this subsection. First, in the multi-point estimates, markers that are far away from the flanking markers will not contribute too much for the improvement of the estimates, since those markers segregate almost independently from the flanking markers and hence contain little information about the sharing of IBD alleles at the flanking markers. Second, if the flanking markers have a relatively large number of alleles, the multi-point estimates will not be much better than the local estimates. This is because that, when the flanking markers have more alleles,

they are more polymorphic and more informative and hence the sharing of IBD alleles can be determined with more certainty.

4.3 A power comparison between the multi-point and the two-point interval mapping

In this section, we examine the type I error rate and evaluate the power of the multi-point interval mapping, and draw a comparison between the multi-point interval mapping and the two-point interval mapping.

A simulation study is conducted to examine the type I error probability of the multi-point interval mapping. We consider the case with 6 successive markers. All markers are equally spaced and have the same number of alleles. For each combination of marker allele number (3 or 6) and interval length (10cM or 20cM), 200 sib pairs are generated. For each sib pair, at each marker, two multi-point estimates of the IBD proportion are calculated, one using the information in itself and all markers to its left ("left hand estimate"), which will be used when this marker acts as a left flanking marker of an interval, the other using the information in itself and all markers to its right ("right hand estimate"), which will be used when this marker acts as a right flanking marker of an interval. Here we should note that, for the leftmost marker, M_1, only the left hand estimate of π_{M_1} need to be calculated since there is no interval to the its left, and its left hand estimate of π_{M_1} equals its local estimate of π_{M_1} since there is no marker to its left.

Similarly only the right hand estimate of π_{M_6} need to be calculated and it is equal to the local estimator of π_{M_6}. The local estimates of the IBD proportion at the markers are also calculated for the two-point interval mapping. Then for each interval, two linear regressions are performed using respectively the multi-point and the local estimates of the IBD proportions at the relevant flanking markers, and the modified Wald statistics are calculated and compared to the corresponding thresholds. The whole procedure is replicated 2000 times, and the proportion of rejections is taken as the simulated type I error probability. The simulated type I error probabilities are reported in Table 4.2.

In Table 4.2, under each combination of marker allele number and interval length, the type I error probability is reported for each of the 5 intervals. The numbers in normal size are calculated with the multi-point estimates of π_Ms, and the bracketed numbers in footnote size are calculated with the local estimates of π_Ms. Though the deviations of the simulated type I error probability from the nominal value are not trivial under some of the situations considered here, for example when $\alpha = 0.05$, the number of alleles is 3 and the interval length is 20cM, the simulated type I error probability for the third interval is 0.065 and 0.067 for the multi-point and the two-point interval mapping respectively, in view of the small sample size and the small number of replications, they are still acceptable.

To look into the power of the multi-point interval mapping and draw a comparison between the multi-point and the two-point interval mapping, a simulation study is conducted.

Table 4.2: Simulated actual levels of the multi-point and two-point interval mapping

No. of alleles	Interval length	Type of estimate	Interval # 1	2	3	4	5
α=0.01							
3	10cM	multi	0.012	0.012	0.011	0.012	0.009
		local	(0.012)	(0.014)	(0.010)	(0.006)	(0.010)
	20cM	multi	0.010	0.009	0.010	0.014	0.015
		local	(0.010)	(0.007)	(0.012)	(0.012)	(0.016)
6	10cM	multi	0.008	0.010	0.008	0.009	0.011
		local	(0.009)	(0.010)	(0.008)	(0.010)	(0.010)
	20cM	multi	0.010	0.014	0.012	0.012	0.012
		local	(0.011)	(0.013)	(0.014)	(0.010)	(0.013)
α=0.05							
3	10cM	multi	0.055	0.058	0.057	0.056	0.045
		local	(0.059)	(0.057)	(0.053)	(0.051)	(0.052)
	20cM	multi	0.051	0.045	0.065	0.063	0.054
		local	(0.053)	(0.053)	(0.067)	(0.059)	(0.054)
6	10cM	multi	0.048	0.051	0.049	0.047	0.050
		local	(0.047)	(0.048)	(0.050)	(0.048)	(0.050)
	20cM	multi	0.049	0.061	0.054	0.058	0.055
		local	(0.049)	(0.057)	(0.056)	(0.059)	(0.054)

A single QTL is assumed to locate in the region spanned by 6 markers which are as described above. The heritability is fixed at 0.5, and the genetic variance is 0.5. The simulated QTL has 3 alleles, whose contributions to the trait value are $\frac{\sqrt{6}}{4}$, 0 and -$\frac{\sqrt{6}}{4}$, respectively. For each combination of marker allele number (3 or 6), QTL location (middle of the first interval or middle of the third interval) and interval length (10cM or 20cM), 200 sib pairs are generated. Then the same procedure as we did for the type I error probability is repeated 2000 times, and the proportion of rejections is taken as the simulated powers and reported in Table 4.3.

It can be seen from Table 4.3 that both interval mapping methods can locate the QTL in the correct interval in terms of the highest power. The estimated power decreases monotonically to the right when the QTL lies in the first interval, and it decreases nearly symmetrically to both sides when the QTL lies in the third interval. These are just as we expected. Next, we will compare the power of the two interval mapping methods. It is found that when the markers are less polymorphic (with 3 alleles only), the multi-point interval mapping method is more powerful than the two-point interval mapping method. The reason is that, such flanking markers are far from completely informative, the QTL information they contain is not enough, and the markers beyond the two flanking ones may contain a large part of the QTL information. This also explains why the multi-point interval mapping method is not more powerful than the two-point interval mapping method when the marker allele number increases to 6. For 6-allele markers, the proportion of being completely informative is as high as 67.13% (Fulker *et al*., 1995). However, if the flanking markers are far away from the QTL, they still

Table 4.3: Simulated powers of the multi-point and two-point interval mapping

Allele number	QTL location	Interval length	Interval 1	Interval 2	Interval 3	Interval 4	Interval 5
α=0.01							
3	1st	10cM	0.305 (0.289)	0.274 (0.209)	0.142 (0.085)	0.072 (0.044)	0.041 (0.030)
		20cM	0.254 (0.240)	0.196 (0.148)	0.061 (0.037)	0.019 (0.021)	0.018 (0.014)
	3rd	10cM	0.171 (0.092)	0.302 (0.219)	0.313 (0.304)	0.293 (0.223)	0.176 (0.095)
		20cM	0.070 (0.041)	0.202 (0.168)	0.266 (0.241)	0.198 (0.159)	0.056 (0.037)
6	1st	10cM	0.329 (0.330)	0.293 (0.257)	0.121 (0.086)	0.063 (0.052)	0.035 (0.031)
		20cM	0.282 (0.281)	0.213 (0.182)	0.053 (0.045)	0.016 (0.018)	0.012 (0.010)
	3rd	10cM	0.132 (0.096)	0.276 (0.252)	0.298 (0.313)	0.290 (0.248)	0.133 (0.098)
		20cM	0.047 (0.041)	0.221 (0.196)	0.300 (0.286)	0.232 (0.194)	0.061 (0.044)
α=0.05							
3	1st	10cM	0.604 (0.571)	0.546 (0.457)	0.369 (0.254)	0.238 (0.164)	0.166 (0.125)
		20cM	0.496 (0.492)	0.414 (0.347)	0.179 (0.131)	0.096 (0.077)	0.074 (0.063)
	3rd	10cM	0.396 (0.261)	0.580 (0.482)	0.615 (0.587)	0.584 (0.471)	0.400 (0.259)
		20cM	0.206 (0.147)	0.435 (0.366)	0.530 (0.511)	0.439 (0.376)	0.192 (0.140)
6	1st	10cM	0.632 (0.632)	0.591 (0.545)	0.344 (0.279)	0.201 (0.172)	0.132 (0.114)
		20cM	0.572 (0.563)	0.469 (0.426)	0.171 (0.137)	0.086 (0.080)	0.068 (0.061)
	3rd	10cM	0.332 (0.276)	0.573 (0.525)	0.611 (0.622)	0.579 (0.536)	0.342 (0.274)
		20cM	0.188 (0.166)	0.468 (0.427)	0.582 (0.573)	0.470 (0.430)	0.183 (0.159)

cannot contain much of the QTL information, and thus the multi-point interval mapping method is a bit superior to the two-point interval mapping method, as we observed for the cases of 6-allele markers and 20cM intervals.

Chapter 5

Likelihood Ratio Test for the Interval Mapping of QTL

Testing the existence of QTL is an important issue for the interval mapping. Lander and Botstein (1989) used the LOD score method, which is equivalent to the likelihood ratio test. They claimed the threshold depended on the size of the genome and the density of genotyped markers, and they gave different threshold formulas for the sparse-map case and the dense-map case and suggested to use extensive numerical simulations to determine thresholds for the intermediate marker densities. They also suggested that, for general cases, the threshold for LOD score was between 2 and 3. Though the performance of their thresholds may be good enough, they did not make clear what the distribution of the LOD statistic was. Haley and Knott (1992) used the likelihood ratio test statistic and approximated its distribution with the classical χ^2_p (p is the number

of parameters) distribution. As we know, the χ^2_p approximation to the distribution of the likelihood ratio statistic is appropriate only when the null parameter space is in the interior of the total parameter space. However, in the interval mapping problem, the null parameter space is on the boundary of the total parameter space. Fulker and Cardon (1994) and Fulker *et al*. (1995) suggested to use the t statistic, $\hat{\beta}/SE(\hat{\beta})$, to test the existence of the QTL, and they claimed that the t statistic approximately followed the t distribution and thus they used the critical values of the standard normal distribution as the thresholds (when the sample size is large enough). As we showed in Chapter 2, the t statistic they used is the maximum among all those t scores at the putative QTL locations, and using normal critical values for this t statistic will cause very high false positiveness.

The likelihood ratio test is the most powerful test theoretically. Though the interval mapping problem does not satisfy a condition for the χ^2_p approximation, the derivation of the asymptotic distribution of the likelihood ratio statistic for the interval mapping is not too prohibitive since the number of parameters is small. Several authors have studied the large sample properties of the distribution of the likelihood ratio statistic under certain nonstandard situations (Chernoff 1954, Self and Liang 1987). Self and Liang (1987) gave examples and illustrated that the asymptotic distribution of the likelihood ratio statistic was either a mixture of chi-square or a mixture of normal under various conditions. Once the mixture structure is clear, the threshold values can be calculated numerically by any statistical package. We are going to apply their method and figure out the asymptotic distribution of the likelihood ratio statistic for the interval mapping.

5.1 Likelihood ratio test for the interval mapping

For the interval mapping model (2.7):

$$Z(\omega) = \alpha + \beta_1\hat{\pi}_1 + \beta_2\hat{\pi}_2 + e,$$

where $e \sim \mathrm{N}(0, \sigma_e^2)$, $\beta_1 = 2\sigma_g^2\beta_{M1}$ (≥ 0) and $\beta_2 = 2\sigma_g^2\beta_{M2}$ (≥ 0), to test whether the QTL exists in the interval is equivalent to test the hypothesis:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = 0.$$

Suppose we take a sample of size n, and denote

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}, \qquad \boldsymbol{\pi} = \begin{pmatrix} \hat{\pi}_{11} & \hat{\pi}_{21} \\ \hat{\pi}_{12} & \hat{\pi}_{22} \\ \vdots & \vdots \\ \hat{\pi}_{1n} & \hat{\pi}_{2n} \end{pmatrix},$$

where $Z_i = Z_i(\omega)$ for simplicity. Since we are only interested in the slopes β_1 and β_2, we can simplify the problem by centralizing the variables first. Let $\tilde{Z}_i = Z_i - \bar{Z}$, $\tilde{\hat{\pi}}_{1i} = \hat{\pi}_{1i} - \bar{\hat{\pi}}_1$ and $\tilde{\hat{\pi}}_{2i} = \hat{\pi}_{1i} - \bar{\hat{\pi}}_2$, and then the model is transformed to

$$\tilde{Z} = \beta_1\tilde{\hat{\pi}}_1 + \beta_2\tilde{\hat{\pi}}_2 + e.$$

The hypothesis remains unchanged, but the number of parameters decreases by 1.

$\boldsymbol{\beta} = (\beta_1, \beta_2)'$ is the parameter of interest, and σ_e^2 is a nuisance parameter. Denote $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma_e^2)'$. Since the regression coefficients β_1 and β_2 are nonnegative, the total parameter space is:

$$\Omega = [0, +\infty)^2 \times (0, +\infty).$$

The parameter space under the null hypothesis H_0 is:

$$\Omega_0 = \{0\}^2 \times (0, +\infty).$$

Given $\tilde{\boldsymbol{\pi}}$, $\tilde{\mathbf{Z}} \sim \mathrm{N}(\tilde{\boldsymbol{\pi}}\boldsymbol{\beta},\ \sigma_e^2\mathbf{I} - \frac{1}{n}\sigma_e^2\mathbf{1}\mathbf{1}')$. When the sample size n is large enough, $\tilde{\mathbf{Z}}$ is asymptotically $\mathrm{N}(\tilde{\boldsymbol{\pi}}\boldsymbol{\beta},\ \sigma_e^2\mathbf{I})$. The likelihood of the centralized data is asymptotically

$$\begin{aligned} L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\tilde{Z}_i) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_e} \exp\left\{-\frac{1}{2\sigma_e^2}(\tilde{Z}_i - \beta_1\tilde{\hat{\pi}}_{1i} - \beta_2\tilde{\hat{\pi}}_{2i})^2\right\} \\ &= (2\pi)^{(-n/2)}(\sigma_e^2)^{(-n/2)} \exp\left\{-\frac{1}{2\sigma_e^2}(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})'(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})\right\}, \quad (5.1) \end{aligned}$$

where n is the sample size. Let $l_n(\boldsymbol{\theta})$ denote the log likelihood function of the sample. Let λ_n be

$$\lambda_n = \frac{\sup\limits_{\Omega_0} L_n(\boldsymbol{\theta})}{\sup\limits_{\Omega} L_n(\boldsymbol{\theta})},$$

and the likelihood ratio statistic is: $-2\ln\lambda_n = -2(\sup_{\Omega_0} l_n(\theta) - \sup_{\Omega} l_n(\theta))$.

Under the assumption of independent and normally distributed random errors, the likelihood ratio test statistic can be written in terms of the residual sum of squares of the full model and the reduced model (Aitkin *et al.*, 1989),

$$\begin{aligned} -2\ln\lambda_n &= n\ln\{RSS_{reduced}/RSS_{full}\} \\ &= n\ln\frac{\inf_{\Omega_0^*}(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})'(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})}{\inf_{\Omega^*}(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})'(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})} \\ &= n\ln\frac{\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}}{\inf_{\Omega^*}(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})'(\tilde{\mathbf{Z}} - \tilde{\boldsymbol{\pi}}\boldsymbol{\beta})} \quad (5.2) \end{aligned}$$

where $\Omega_0^* = \{0\}^2$ and $\Omega^* = [0, +\infty)^2$ are the null and total parameter space of $\boldsymbol{\beta}$. The likelihood ratio statistic represented by formula (5.2) can be computed by numerical methods.

Since Ω_0 is on the boundary of Ω, $-2\ln\lambda_n$ will not follow the classic χ^2_2 distribution, and we need to explore the asymptotic properties of its distribution.

In the next section, the asymptotic distribution of $-2\ln\lambda_n$ will be derived using Taylor series expansion method of Chernoff (1954) and Self and Liang (1987), and we also provide another way of reasoning specific to the normal likelihood function which leads to the same result.

5.2 Deriving the asymptotic distribution of the likelihood ratio statistic

Chernoff (1954) and Self and Liang (1987) applied Taylor series expansion to the log likelihood function, and reduced the distribution of the likelihood ratio statistic to the form of a mixture of chi-square or mixture of normal under certain nonstandard conditions. These results are general for any form of $L_n(\boldsymbol{\theta})$ and thus can be applied here. We will demonstrate their approach with the interval mapping problem as an example.

The following regularity conditions are assumed, which are necessary for the likelihood function f and the maximum likelihood estimates to have certain desirable properties.

1. For almost all Z, the derivatives

$$\frac{\partial \log f}{\partial \theta_i}, \qquad \frac{\partial^2 \log f}{\partial \theta_i\, \partial \theta_j}, \qquad \frac{\partial^3 \log f}{\partial \theta_i\, \partial \theta_j\, \partial \theta_m}$$

exist for every $\boldsymbol{\theta}$ in the closure of a neighborhood of $\boldsymbol{\theta}_0$, N, where θ_0 is the true parameter value.

2. For $\boldsymbol{\theta} \in N$,

$$\left|\frac{\partial f}{\partial \theta_i}\right| < F(Z), \qquad \left|\frac{\partial^2 f}{\partial \theta_i \, \partial \theta_j}\right| < F(Z), \qquad \left|\frac{\partial^3 \log f}{\partial \theta_i \, \partial \theta_j \, \partial \theta_m}\right| < H(Z),$$

where F is finitely integrable and $E\{H(Z)\} < M$, with M independent of $\boldsymbol{\theta}$.

3. For $\boldsymbol{\theta} \in N$, $\left\| E\left\{\frac{\partial \log f}{\partial \theta_i} \; \frac{\partial \log f}{\partial \theta_j}\right\}\right\|$ is finite and positive definite.

For large samples, the likelihood function $L_n(\boldsymbol{\theta})$ is given by formula (5.1). It follows from the above regularity conditions that, for $\boldsymbol{\theta} \in N$,

$$\frac{1}{n} l_n(\boldsymbol{\theta}) = \frac{1}{n} l_n(\boldsymbol{\theta}_0) + \frac{1}{n} U_n'(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \frac{1}{2n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' J_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^3 O_p(1), \quad (5.3)$$

where $U_n(\boldsymbol{\theta}_0)$ is the first order derivative of $l_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$, and $J_n(\boldsymbol{\theta}_0)$ is the negative second order derivative of $l_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$. We can replace $\frac{1}{n} J_n(\boldsymbol{\theta}_0)$ with its expectation $\tilde{J}(\boldsymbol{\theta}_0)$, where

$$\tilde{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{n\sigma_e^2} \tilde{\boldsymbol{\pi}}' \tilde{\boldsymbol{\pi}} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_e^4} \end{pmatrix}. \quad (5.4)$$

For simplicity, we will denote $\tilde{J}(\boldsymbol{\theta}_0)$ with $\tilde{J}$, and $U_n(\boldsymbol{\theta}_0)$ with U for short, and omit the term $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^3 O_p(1)$ from now onwards.

Let $\boldsymbol{\eta} = \boldsymbol{\theta} - \frac{1}{n}\tilde{J}^{-1}U$, and then equation (5.3) becomes:

$$\begin{aligned}
\frac{1}{n}l_n(\boldsymbol{\theta}) &= \frac{1}{n}l_n(\boldsymbol{\theta}_0) + \frac{1}{n^2}U'\tilde{J}^{-1}U + \frac{1}{n}U'(\boldsymbol{\eta}-\boldsymbol{\theta}_0) - \frac{1}{2n^2}U'\tilde{J}^{-1}U - \frac{1}{n}U'(\boldsymbol{\eta}-\boldsymbol{\theta}_0) \\
&\quad -\frac{1}{2}(\boldsymbol{\eta}-\boldsymbol{\theta}_0)'\tilde{J}(\boldsymbol{\eta}-\boldsymbol{\theta}_0) \\
&= \frac{1}{n}l_n(\boldsymbol{\theta}_0) + \frac{1}{2n^2}U'\tilde{J}^{-1}U - \frac{1}{2}(\boldsymbol{\eta}-\boldsymbol{\theta}_0)'\tilde{J}(\boldsymbol{\eta}-\boldsymbol{\theta}_0) \\
&= \frac{1}{n}l_n(\boldsymbol{\theta}_0) + \frac{1}{2n^2}U'\tilde{J}^{-1}U - \frac{1}{2}(\boldsymbol{\gamma}-(\boldsymbol{\theta}-\boldsymbol{\theta}_0))'\tilde{J}(\boldsymbol{\gamma}-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)), \qquad (5.5)
\end{aligned}$$

where $\boldsymbol{\gamma} = \frac{1}{n}\tilde{J}^{-1}U$, and $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \frac{1}{n}\tilde{J}^{-1})$ since $\mathrm{Var}(U) = n\tilde{J}$. Therefore,

$$\begin{aligned}
-2\ln\lambda_n &= -2(\sup_{\Omega_0} l_n(\theta) - \sup_{\Omega} l_n(\theta)) \\
&\approx \inf_{\Omega_0} n(\boldsymbol{\gamma}-(\boldsymbol{\theta}-\boldsymbol{\theta}_0))'\tilde{J}(\boldsymbol{\gamma}-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)) - \inf_{\Omega} n(\boldsymbol{\gamma}-(\boldsymbol{\theta}-\boldsymbol{\theta}_0))'\tilde{J}(\boldsymbol{\gamma}-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)) \\
&= \inf_{C_0}(\boldsymbol{\gamma}-\boldsymbol{\theta})'\tilde{J}(\boldsymbol{\gamma}-\boldsymbol{\theta}) - \inf_{C}(\boldsymbol{\gamma}-\boldsymbol{\theta})'\tilde{J}(\boldsymbol{\gamma}-\boldsymbol{\theta}), \qquad (5.6)
\end{aligned}$$

where $C_0 = \sqrt{n}(\Omega_0 - \boldsymbol{\theta}_0)$, $C = \sqrt{n}(\Omega - \boldsymbol{\theta}_0)$, and $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \tilde{J}^{-1})$. If the null hypothesis is true, $\boldsymbol{\theta}_0 \in \Omega_0$, we have $C_0 = \{0\}^2 \times R$ and $C = [0, +\infty)^2 \times R$ as n tends to infinity.

From equation (5.4), we have

$$(\boldsymbol{\gamma}-\boldsymbol{\theta})'\tilde{J}(\boldsymbol{\gamma}-\boldsymbol{\theta}) = \frac{1}{2\sigma_e^4}(\gamma_3-\theta_3)^2 + (\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})'\frac{1}{n\sigma_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta}),$$

where $\tilde{\boldsymbol{\gamma}} = (\gamma_1, \gamma_2)'$, and $\tilde{\boldsymbol{\gamma}} \sim \mathrm{N}(\mathbf{0}, n\sigma_e^2(\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}})^{-1})$. Since θ_3 can take any real number in C_0 and C, hence

$$\inf(\boldsymbol{\gamma}-\boldsymbol{\theta})'\tilde{J}(\boldsymbol{\gamma}-\boldsymbol{\theta}) = \inf(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})'\frac{1}{n\sigma_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta}),$$

with $\theta_3 = \gamma_3$. The corresponding space for $\boldsymbol{\beta}$ is $\tilde{C}_0 = \{0\}^2$ and $\tilde{C} = [0, +\infty)^2$.

Now let $\boldsymbol{J} = (n\hat{\sigma}_e^2)^{-1}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}$, and the spectral decomposition of $\boldsymbol{J}$ is $P\Lambda P'$, P is an orthogonal matrix whose columns are the eigen vectors of $\boldsymbol{J}$ and Λ is a diagonal matrix

whose entries are the eigen values of $\boldsymbol{J}$. Then we have

$$\begin{aligned}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})'\frac{1}{n\sigma_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta}) &= (\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})'P\Lambda P'(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta}) = \|\Lambda^{\frac{1}{2}}P'(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})\|^2 \\ &= \|\mathbf{X}-\boldsymbol{\upsilon}\|^2,\end{aligned}$$

where $\mathbf{X} = \Lambda^{\frac{1}{2}}P'\tilde{\boldsymbol{\gamma}}$, $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\upsilon} = \Lambda^{\frac{1}{2}}P'\boldsymbol{\beta}$. $\boldsymbol{\upsilon} \in \mathcal{P}_0 = \{0\}^2$ under the null hypothesis and the total parameter space of $\boldsymbol{\upsilon}$ is $\mathcal{P} = \Lambda^{\frac{1}{2}}P'[0, +\infty)^2$.

Thus the asymptotic representation of the likelihood ratio statistic $-2\ln\lambda_n$ is

$$\inf_{\mathcal{P}_0}\|\mathbf{X}-\boldsymbol{\upsilon}\|^2 - \inf_{\mathcal{P}}\|\mathbf{X}-\boldsymbol{\upsilon}\|^2. \tag{5.7}$$

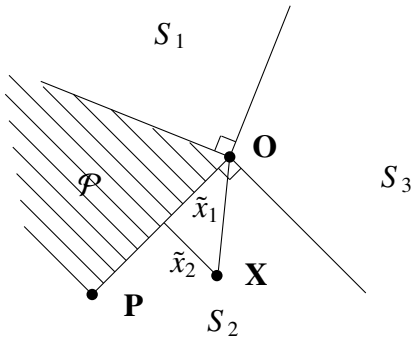


Figure 5.1: Diagram of the parameter space

The distribution of the asymptotic representation of $-2\ln\lambda_n$ (formula (5.7)) will change as $\mathbf{X}$ changes, and this can be illustrated with an example. In figure 5.1, the shaded region represents the total parameter space $\mathcal{P}$, the origin $\mathbf{O}$ is the null parameter space $\mathcal{P}_0$. The value and distribution of formula (5.7) for $\mathbf{X}$ in different region is as

follows:

$$\begin{cases} \mathcal{P}: & X_1^2 + X_2^2 & \chi_2^2 \\ S_1 \& S_2: & \tilde{x}_1^2 & \chi_1^2 \\ S_3: & 0 & \chi_0^2 \end{cases},$$

where $\tilde{x}_1$ is the projection of **XO** on the boundary of region $\mathcal{P}$ and S_2 when **X** is in S_2. Note that, $\tilde{x}_1$ and $\tilde{x}_2$ are the new coordinates of **X** after a rotation of the old axes that makes **OP** on one of the new axes, and the rotation of the axes equals an orthogonal transformation of **X**, and hence $\tilde{x}_1$ is also a standard normal random variable. Furthermore, the asymptotic distribution of the likelihood ratio statistic $-2\ln\lambda_n$ has a mixture structure:

$$(0.5 - r)\chi_0^2 + 0.5\chi_1^2 + r\chi_2^2.$$

The mixing probability r is the proportion of the shaded region $\mathcal{P}$ in the whole 2-dimensional space, and it can be calculated as

$$r = \cos^{-1}\left\{\frac{(1\ 0)P\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}P'(0\ 1)'}{||(\Lambda^{\frac{1}{2}}P'(1\ 0)')||||(\Lambda^{\frac{1}{2}}P'(0\ 1)')||}\right\}\Big/2\pi = \cos^{-1}\left\{\frac{\boldsymbol{J}_{12}}{\sqrt{\boldsymbol{J}_{11}\boldsymbol{J}_{22}}}\right\}\Big/2\pi,$$

and the asymptotic threshold values for $-2\ln\lambda_n$ corresponding to each possible mixing probability r can be pre-calculated by statistical softwares.

In the following part till the end of this section, we will introduce another approach to deriving the asymptotic distribution of the likelihood ratio statistic specific to the normal likelihood function.

Denote the least squares estimate of $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}})^{-1}\tilde{\boldsymbol{\pi}}'\tilde{\mathbf{Z}}, \quad \text{and} \quad \hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}_0, \sigma_e^2(\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}})^{-1}).$$

Then $\hat{\sigma}_e^2$ can be estimated by $\frac{1}{n}(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\hat{\boldsymbol{\beta}})'(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\hat{\boldsymbol{\beta}})$, and we can decompose $(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\boldsymbol{\beta})'(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\boldsymbol{\beta})$ as:

$$\begin{aligned}(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\boldsymbol{\beta})'(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\boldsymbol{\beta}) &= (\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\hat{\boldsymbol{\beta}})'(\tilde{\mathbf{Z}}-\tilde{\boldsymbol{\pi}}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= n\hat{\sigma}_e^2 + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= n\hat{\sigma}_e^2\left(1 + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right).\end{aligned}$$

Then we have:

$$-2\ln\lambda_n = n\ln\left(\frac{1+\inf_{\Omega_0^*}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{1+\inf_{\Omega^*}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}\right).$$

It can be verified that

$$\inf_{\Omega_0^*}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{n\to\infty} 0 \quad \text{and} \quad \inf_{\Omega^*}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{n\to\infty} 0.$$

Finally we have

$$\begin{aligned}-2\ln\lambda_n &\approx n\left(\inf_{\Omega_0^*}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \inf_{\Omega^*}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right) \\ &= \inf_{\Omega_0^*}(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) - \sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) - \sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)) \\ &\quad - \inf_{\Omega^*}(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) - \sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) - \sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)) \\ &= \inf_{\tilde{\mathbf{C}}_0}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta}) - \inf_{\tilde{\mathbf{C}}}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta})'\frac{1}{n\hat{\sigma}_e^2}\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\beta}),\end{aligned}$$

where $\tilde{\boldsymbol{\gamma}} = \sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \sim \mathrm{N}(\mathbf{0}, n\hat{\sigma}_e^2(\tilde{\boldsymbol{\pi}}'\tilde{\boldsymbol{\pi}})^{-1})$, and $\tilde{\mathbf{C}}_0 = \sqrt{n}(\Omega_0^*-\boldsymbol{\beta}_0) = \{0\}^2$, $\tilde{\mathbf{C}} = \sqrt{n}(\Omega^*-\boldsymbol{\beta}_0) = [0,+\infty)^2$ if the true value $\boldsymbol{\beta}_0$ is in Ω_0^*.

This asymptotic representation of the likelihood ratio statistic is exactly equal to formula (5.7). The next steps for deriving the asymptotic distribution are the same as above, and will not be repeated here.

5.3 Simulation studies

Simulation studies were conducted to compare the likelihood ratio test with the modified Wald test proposed in Chapter 2.

We consider the case with 6 equally spaced successive markers that have the same number of alleles. A single QTL is assumed to locate half way between the third and the fourth marker. The heritability is fixed at 0.5, and the genetic variance is 0.5. The simulated QTL has only 3 alleles, whose contributions to the trait value are $\frac{\sqrt{6}}{4}$, 0 and $-\frac{\sqrt{6}}{4}$, respectively. For each combination of marker allele number (2, 4 or 8) and interval length (10cM or 20cM), 200 sib pairs are simulated. For each sib pair, at each marker, the local estimate and the multi-point estimate of the IBD proportion are obtained. Next, for each interval, the likelihood ratio statistic and the modified Wald statistic using both local and multi-point estimates of the IBD proportion are calculated and compared to their corresponding threshold values. This procedure is replicated 2,000 times, and the proportion of rejections is taken as the simulated power and shown in figure 5.2 and figure 5.3.

Figure 5.2 shows the power graphs of the 4 tests: multi-point interval mapping model using modified Wald test (**mulWd**), multi-point interval mapping model using likelihood ratio test (**mullLR**), two-point interval mapping model using modified Wald test (**twoWd**) and two-point interval mapping model using likelihood ratio test (**twoLR**). The nominal level of the tests presented in Figure 5.2 is 0.01. Each panel of the figure corresponds to a particular combination of interval length and marker allele

number. Figure 5.3 shows the power graphs under the same settings but nominal level 0.05.

It can be seen from these 2 figures that, under each setting, the power graphs of the 4 tests are all inverse U-shaped and they all can locate the putative QTL in the correct interval in terms of the highest power among the 5 intervals, as we expected.

We now look into the effect of marker allele number on the power of the *mulLR* test and the *twoLR* test. It can be seen that the power of the *twoLR* test is always lower than that of the *mulLR* test, but as the marker allele number increases the difference decreases obviously. The reason was given in Chapter 4: markers with more alleles are more informative and thus their local estimates of π_M tend to be as good as the multi-point estimates. It also can be observed that the powers of the two likelihood ratio tests keep increasing as the marker allele number increases. This is because that, more polymorphic markers are more informative and thus the tests become more powerful.

The effect of interval length on the powers of the two likelihood ratio tests is similar to that of marker allele number. The power of the *mulLR* test is always higher than that of the *twoLR* test, and the difference decreases as the interval length increases. The explanation was also given in Chapter 4: as two markers get further away from each other, the recombination fraction between them will get closer to 0.5, they tend to behave more independently, the estimate of π_M at one marker will depend less on its nearby markers, and thus the multi-point estimate of π_M is less superior to the local estimate. Different from the effect of the marker allele number, increasing the interval length decreases the

powers of the two likelihood ratio tests since when the distance between the flanking markers gets longer, the flanking markers will contain less information on the QTL, and thus the tests become less powerful.

To evaluate the efficiency of the likelihood ratio test and the modified Wald test, extensive comparisons are drawn between the *mulLR* test and the *mulWd* test and between the *twoLR* test and the *twoWd* test. The results of the two comparisons are completely the same in all aspects. The likelihood ratio test is more powerful than the modified Wald test under every combination of parameters. The difference in power increases as the marker allele number increases. This indicates that the likelihood ratio test can absorb marker information more effectively. However the difference in power decreases as the interval length increases. Since when the intervals are longer, the markers will contain less information on the QTL, and thus the likelihood ratio test cannot absorb more marker information than the modified Wald test as it does in the case of shorter interval. Furthermore, we can infer that the likelihood ratio test is most beneficial for short intervals and highly polymorphic markers, whereas the modified Wald test is comparable to the likelihood ratio test only when the markers are sparse and far from completely polymorphic. The above inferences are pretty well demonstrated by the results of our simulation studies.

Figure 5.2: Power comparison between the LR test and the modified Wald test for multi-point and two-point interval mapping ($\alpha = 0.01$)

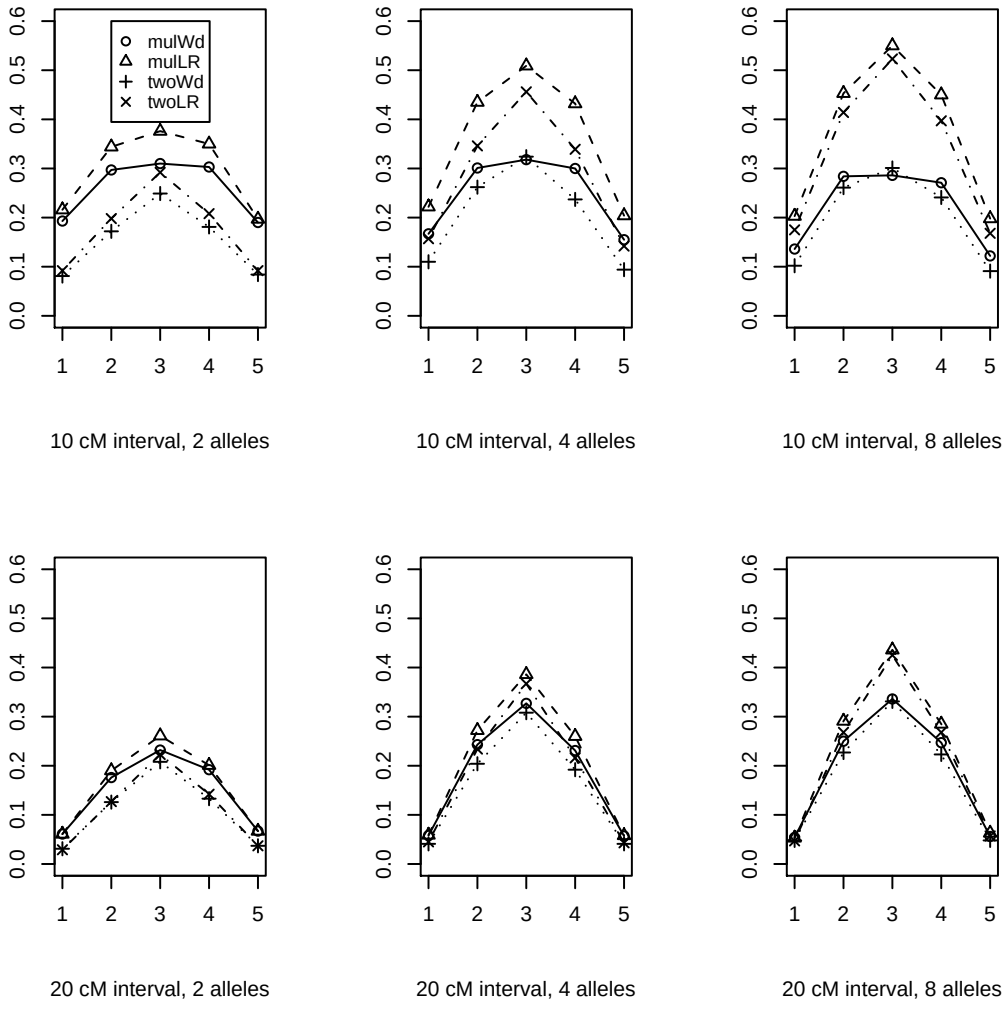
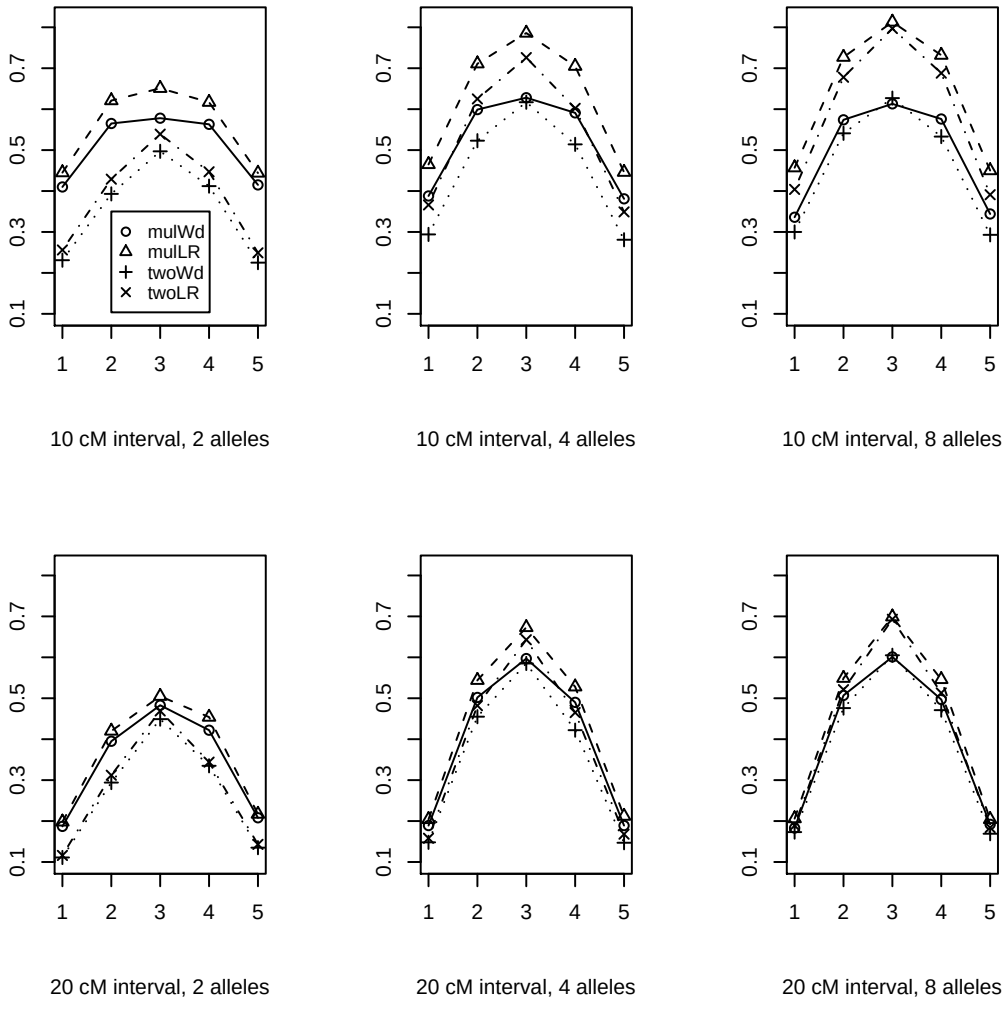


Figure 5.3: Power comparison between the LR test and the modified Wald test for multi-point and two-point interval mapping ($\alpha = 0.05$)



Chapter 6

Conclusion and Further Research

6.1 Conclusion

It has been shown in Fulker and Cardon (1994) that the interval mapping approach is more powerful in detecting QTL than single marker mapping methods and that it provides a more precise estimate of the QTL location. The interval mapping approach is especially beneficial when the markers are relatively coarse. However, since the type I error probability is not appropriately controlled by the nominal t-test, in fact, the type I error probability is inflated, the nominal t-test could lead to undesirable false positiveness in QTL mapping. The modified Wald test developed in this thesis effectively removes this pitfall. It makes the more powerful interval mapping approach more reliable for QTL mapping in human beings.

In real QTL mapping problems, the genome-wide search is more appropriate than the single interval mapping. However, the multiple tests associating with the genome-wide search will inevitably inflate the overall type I error probability. In this thesis, we propose a genome-wide search strategy using the modified Wald statistic given in Chapter 2, and we also provide an approach to simulating the unified thresholds. Simulation studies show that the unified thresholds are able to control the overall type I error probability. Simulation results also suggest that the power of the genome-wide search is affected simultaneously by the interval length, the genetic variance, and the relative distance between QTLs if there are more than one QTL.

The interval mapping method only makes use of the two flanking markers. However, when the two flanking markers are not completely informative, only a part of the QTL information is contained in the flanking markers, and the rest is contained in some nearby markers. In this thesis, we formulate a new model for the multi-point interval mapping, in which the IBD proportions at the flanking markers are estimated with the joint distribution of the numbers of alleles IBD at multiple markers. Simulation results show that the type I error probability of the multi-point interval mapping matches the nominal value well. A comparison between the multi-point interval mapping and the two-point interval mapping shows that, the multi-point interval mapping is more powerful than the two-point interval mapping if the flanking markers are less polymorphic (<6 alleles), but it is not the case for the more polymorphic markers since the two flanking markers have already contained much enough of the QTL information. The simulation results also show that, for the more polymorphic markers, the multi-point interval map-

ping is superior to the two-point interval mapping only if the intervals are very long (>20cM), because the two flanking markers cannot carry much of the QTL information when they are far from the QTL.

The likelihood ratio test is the most powerful test theoretically. However, the interval mapping problem is not a standard situation for the χ^2_p approximation of the LR statistic. In this thesis, we apply the results of Self and Liang (1987) on the asymptotic properties of the LR statistic under non-standard conditions, and deduce that the asymptotic distribution of the LR statistic is a mixture of χ^2_1 and χ^2_2. Simulation results show that, the LR test is always more powerful than the modified Wald test, the power of the LR test increases as the marker allele number increases, and it decreases as the interval length increases. Furthermore, we can infer that the likelihood ratio test is most beneficial for short intervals and highly polymorphic markers.

6.2 Topics for further research

The variance components methods are more powerful than the Haseman-Elston regression methods in human QTL mapping if the QT is normally distributed or nearly so. However, the variance components methods cannot provide QTL location estimates. The interval mapping methods for human QTL can detect the existence of QTL and estimate its location if it exists. Though interval mapping has been proved to be more powerful than single marker mapping, it is still regression based. If we combine the

variance components model with the interval mapping idea, the power of detecting the QTL is expected to be improved. If only sib pair data are used, to extend interval mapping to variance components interval mapping, we just need to formulate the variance of $(\beta_1\hat{\pi}_{M1} + \beta_2\hat{\pi}_{M2})$ for each sib pair. If pedigree data are used, more amendments are needed. We may need to formulate the variance-covariance structure of $(\beta_1\hat{\pi}_{M1} + \beta_2\hat{\pi}_{M2})$ for all relative pairs in the same pedigree.

The likelihood ratio test for the interval mapping of single QTL is shown to be very powerful in this thesis. However, most of the quantitative traits in the nature are genetically controlled by more than one QTL. Therefore, it is necessary to extend the likelihood ratio test to multiple QTL cases. The asymptotic representation of the LR statistic for the multiple QTL case is the same as that for the single QTL case (formula 5.7), but the derivation of its distribution, or the distance-minimization process, becomes much more complicated due to high dimension of the total parameter space. We can consider some numerical methods for simulating the critical values if the asymptotic distribution of the LR statistic is too hard to derive.

In the unified interval mapping regression model 2.7, the random error e_i is assumed to follow $\mathrm{N}(0, \sigma_e^2)$. However, this assumption may be incorrect. When the QT is normal or nearly normal, e_i will follow certain χ^2 distribution (central or noncentral). The QT can also be non-normal, and the distribution of e_i will be more complicated. Therefore, we should not restrict ourselves to the simple linear regression, which relies completely on the normality assumption. The generalized linear models are more appropriate in

practice. The generalized linear models can provide a more accurate estimate of the QTL location, and their goodness of fit can be better than that of the simple linear regression model.

References

Aitkin, M., Anderson, D., Francis, B., Hinde, J. (1989). *Statistical Modelling in GLIM*. Oxford University Press, Oxford.

Almasy, L., Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics* **62**, 1198–1211.

Almasy, L., Dyer, T. D., Blangero, J. (1997). Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genetic Epidemiology* **14**, 953–958.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *The American Journal of Human Genetics* **54**, 535–543.

Beckmann, J. S., Soller, M. (1988). Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. *Theoretical and Applied Genetics* **76**, 228–236.

Botstein, D., White, R. L., Skolnick, M. Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *The American Journal of Human Genetics* **32**, 314–331.

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal* **30**, 44–52.

Broman, K. W., Speed, T. P. (1999). A review of methods for identifying QTLs in experimental crosses. In: Seillier-Moiseiwitsch, F., ed., Statistics in Molecular

Biology and Genetics. Vol. 33 of IMS Lecture Notes–Monograph Series, 114–142.

Campbell, M. A., Elston, R. C. (1971). Relatives of probands: models for preliminary genetic analysis. *Annals of Human Genetics* **35**, 225–236.

Carlborg, O., Andersson, L., Kinghorn, B. (2000). The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics* **155**, 2003–2010.

Chen, Z., Chen, H. (2005). On some statistical aspects of the interval mapping for QTL detection. *Statistica Sinica* **15**, 909–925.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25**, 573–578.

Churchill, G. A., Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* **3**, 43–52.

Doerge, R. W., Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Doerge, R. W., Weir, B. S., Zeng, Z-B. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12**, 195–219 .

Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23**, 34–63.

Drigalenko, E. (1998). How sib pairs reveal linkage. *The American Journal of Human Genetics* **63**, 1242–1245.

Edwards, M. D., Stuber, C. W., Wendel, J. F. (1987). Molecular marker- facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* **116**, 113–125.

Elston, R. C., Buxbaum, S., Jacobs, K. B., Olson, J. M. (2000). Haseman and Elston revisited. *Genetic Epidemiology* **19**, 1–17.

Elston, R. C., Keats, B. J. B. (1985). Genetic analysis workshop III: Sib pair analyses to determine linkage groups and to order loci. *Genetic Epidemiology* **2**, 211–213.

Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *American Journal of Human Genetics* **71**, 217–222.

Feingold, E., Brown, P. O., Siegmund, D. (1993). Gaussian models for genetic linkage

analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics* **53**, 234–251.

Forrest, W. (2001). Weighting improves the "new Haseman-Elston" method. *Human Heredity* **52**, 47–54.

Fulker, D. W., Cardon, L. R. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics* **54**, 1092–1103.

Fulker, D. W., Cherny, S. S., Cardon, L. R. (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *The American Journal of Human Genetics* **56**, 1224–1233.

Ghosh, S., Begleiter, H., Porjesz, B., Chorlian, D. B., Edenberg, H. J., Foroud, T., Goate, A., Reich, T. (2003). Linkage mapping of beta 2 EEG waves via non-parametric regression. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics* **118**, 66–71.

Ghosh, S., Majumder, P. P. (2000). A two-stage variable-stringency semiparametric method for mapping quantitative trait loci with the use of genomewide scan data on sib-pairs. *The American Journal of Human Genetics* **66**, 1046–1061.

Haley, C. S., Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Haley, C. S., Knott, S. A., Elston, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.

Haseman, J. K., Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.

Hoeschele, I., VanRaden, P. M. (1993). Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. *Theoretical and Applied Genetics* **85**, 953–960.

Jansen, R. C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jansen, R. C., Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Kao, C-H., Zeng, Z-B. (1997). General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653–665.

Kao, C-H., Zeng, Z-B., Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Knapp, S. J. (1991). Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. *Theoretical and Applied Genetics* **81**, 333–338

Kruglyak, L., Daly, M. J., Lander, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *American Journal of Human Genetics*, **56**, 519–527.

Kruglyak, L., Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

Lander, E. S., Botstein, D. (1986). Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 7353–7357.

Lander, E. S., Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lander, E. S., Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

Lange, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis.* Springer, New York.

Li, C. C., Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**, 347–360.

Liu, B-H. (1997). *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. CRC press.

Luo, Z. W., Kearsey, M. J. (1989). Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. *Heredity* **63**, 401–408.

Lynch, M., Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, Massachusetts.

Majumder, P. P., Ghosh, S. (2005). Mapping quantitative trait loci in humans: Achievements and limitations. *The Journal of Clinical Investigation* **115**, 1419–1424.

Mitchell, B. D., Ghosh, S., Schneider, J. L., Birznicks, G., Blangero, J. (1997). Power of variance component linkage analysis to detect epistasis. *Genetic Epidemiology* **14**, 1017–1022.

Olson, J. M., Wijsman, E. M. (1993). Linkage between quantitative trait and marker loci: Methods using all relative pairs. *Genetic Epidemiology* **10**, 87–102.

Piepho, H-P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**, 425–432.

Putter, H., Sandkuijl, L. A., van Houwelingen, J. C. (2002). Score test for detecting linkage to quantitative traits. *Genetic Epidemiology* **22**, 345–355.

Rebai, A., Goffinet, B., Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235–240.

Rebai, A., Goffinet, B., Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics* **51**, 87–99.

SAGE (1989). Statistical analysis for genetic epidemiology, release 2.4 01. Department of Biometry and Genetics, LSU Medical Center, New Orleans.

Satagopan, J. M., Yandell, B. S., Newton, M. A., Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. *Genetics* **144**, 805–816.

Self, S. G., Liang, K-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.

Sham, P. C., Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analysis for sib pairs. *The American Journal of Human Genetics* **68**, 1527–1532.

Sham, P. C., Purcell, S., Cherny, S. S., Abecasis, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics* **71**, 238–253.

Sillanp*ää*, M. J., Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**, 1605–1619.

Simpson, S. P. (1989). Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* **77**, 815–819.

Simpson, S. P. (1992). Correction: Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* **85**, 110–111.

Soller, M., Brody, T., Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

Stern, M. P., Duggirala, R., Mitchell, B. D., Reinhart, L. J., Shivakumar, S., Shipman, P. A., Uresandi, O. C., Benavides, E., Blangero J., O'Connell P. (1996). Evidence for linkage of regions on chromosomes 6 and 11 to plasma glucose concentrations in Mexican Americans. *Genome Research* **6**, 724–734.

Szatkiewicz, J. P., Cuenco, K. T., Feingold, E. (2003). Recent advances in human quantitative-trait-locus mapping: comparison of methods for discordant sibling pairs. *The American Journal of Human Genetics* **73**, 874–885.

Tang, H-K., Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2**, 147–162.

Thoday, J. M. (1961). Location of polygenes. *Nature* **191**, 368–370.

Towne, B., Siervogel, R. M., Blangero, J. (1997). Effects of genotype-by-sex interaction on quantitative trait linkage analysis. *Genetic Epidemiology* **14**, 1053–1058.

Uimari, P., Hoeschele, I. (1997). Mapping-Linked quantitative trait loci using Bayesian analysis and Markov Chain Monte Carlo algorithms. *Genetics* **146**, 735–743.

Visscher, P. M., Hopper, J. L. (2001). Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Annals of Human Genetics* **65**, 583–601.

Wang, K., Huang, J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *The American Journal of Human Genetics* **70**, 412–424.

Wright, F. A. (1997). The phenotypic difference discards sib-pair QTL linkage information. *The American Journal of Human Genetics* **60**, 740–742.

Xu, X., Weiss, S., Xu, X., Wei, L. J. (2000). A unified Haseman-Elston method for testing linkage with quantitative traits. *The American Journal of Human Genetics* **67**, 1025–1028.

Zeng, Z-B. (1993). Theoretical basis for seperation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10972–10976.

Zeng, Z-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Zeng, Z-B., Kao, C-H., Basten, C. J. (1999). Estimating the genetic architecture of quantitative traits. *Genetical Research* **74**, 279–289.

Zou, F., Fine, J. P., Hu, J., Lin, D. Y. (2004). An efficient resampling method for

assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* **168**, 2307–2316.